Journal of Applied Psychology
1992, Vol. 77, No. 3, 298-308

0021-9010/92/$3.00

Equivalence of Constructs Measured by Job-Specific and Commercially Available Aptitude Tests

Keith Hattrup, Neal Schmitt, and Ronald S. Landis
Michigan State University

This study examined the equivalence of constructs underlying scores on tests designed to measure the knowledge, skills, and abilities required by specific jobs with the constructs underlying scores on aptitude tests taken from published test batteries. Several models of construct equivalence, differing in their assumptions about factor patterns, factor loadings, and variable uniquenesses, were assessed with confirmatory factor analysis. Results indicated that the job-specific tests measured constructs that were essentially equivalent with the constructs measured by the commercially available tests, although the magnitude of unique residual variances differed among the two sets of tests. Furthermore, multiple-groups confirmatory factor analysis indicated that tests loaded equivalently on shared constructs across several sex and race subgroups, although unique residual variances differed across groups. Practical and theoretical implications are discussed.

Choices of predictor measures in personnel selection contexts typically require decisions regarding whether tests of general aptitudes relevant to success across diverse jobs, or tests designed to measure the specific knowledge, skills, and abilities (KSAs) needed to perform the jobs of interest, should be used. Tests of the former type are commercially available and may be cost effective in the short term. Moreover, substantial meta-analytic evidence suggests that general aptitude tests predict success across a wide variety of jobs (Ghiselli, 1966, 1973; Hunter & Hunter, 1984). General aptitude tests can be criticized, however, for containing items that have no apparent relevance to success for any particular job (Kleiman & Faley, 1985). Determination of the desired predictors for a job requires that the user base his or her decisions on some logical, empirical, or theoretical foundation (Society for Industrial and Organizational Psychology, 1987). Hence, general aptitude tests may come under particularly close scrutiny if a rationale for their choice has not been made explicit (Kleiman & Faley, 1985). Defending the use of tests lacking face validity may also be problematic with judges who have little appreciation for statistical arguments regarding the tests' criterion-related validity (Schmitt & Klimoski, 1991).

Developing new tests to measure the KSAs needed in the performance of specific jobs entails a larger initial investment and a potential requirement for a local criterion-related validation study. Many benefits may accrue, however, from efforts to develop tests designed to measure specific job requirements. Test takers may respond more favorably to such tests and may be more motivated to perform than they are on tests lacking face validity (Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977; Smither & Pearlman, 1991). Because initial testing typically occurs during the applicant's first encounters with the organization, tests that contain items of obvious relevance to the job may communicate greater care and commitment to human resources than tests that do not seem job relevant (Schmitt & Klimoski, 1991). Furthermore, Wernimont and Campbell (1968) pointed out that tests that serve more as samples of job behavior than as signs of future behavior may guard against faking, unfairness of content, and concerns over invasion of privacy.

The development of items reflecting job tasks may also provide partial evidence of the test's content validity. Such test-development efforts, when properly based on the results of a thorough job analysis, substantially reduce the employer's burden of proving the rationale behind the choice of predictors (Kleiman & Faley, 1985; Schmitt & Klimoski, 1991). Nevertheless, one of the greatest potential deficiencies of such job-relevant tests may be a lack of sufficient understanding of the underlying constructs measured by test items. Examination of the constructs underlying scores on tests of job-specific KSAs may reveal that measures that appear unrelated in content or format are in fact redundant in measuring a singular construct (cf. Tenopyr, 1977). Furthermore, although content-oriented test construction may represent an attempt to provide a defensible sampling of the job, the scores derived from such measures may be contaminated by extraneous constructs, such as visual acuity, reading ability, anxiety, noise, and so on (Guion, 1978). Although general aptitude tests may suffer from similar contaminating or redundant variance, job-specific measures may not be developed with as careful attention to the meaning of test scores in relation to the nomological network of constructs contributing to test variance (Messick, 1988). Furthermore, less evidence of the test's performance and potential bias in diverse samples of applicants is typically available with specially constructed measures.

We gratefully acknowledge three anonymous reviewers for their helpful and informative comments on an earlier draft of this article.

The action editor for this article was William C. Howell, who assigned the reviewers and conducted the editorial process indepently of the Editor.

Correspondence concerning this article should be addressed to Keith Hattrup, Department of Psychology, 135 Snyder Hall, Michigan State University, East Lansing, Michigan 48824-1117.

As pointed out by Tenopyr (1977), all tests fundamentally provide measures of underlying constructs, whether the tests are general aptitude tests or pure work samples. Constructs refer to meanings and labels attached to patterns of observable behaviors and phenomena (Binning & Barrett, 1989; Nunnally, 1978). Hence, scores on a test designed to simulate actual performance requirements related to reading dials and gauges on the job may reflect the influence of some underlying construct, such as symbolic interpretation. From this perspective, variance in scores is assumed to arise from variance in an underlying construct, and scores on tests of very different content and format may nevertheless covary to the extent that they share the same constructs (Linn & Werts, 1979; Tenopyr, 1977; Turban, Sanders, Francis, & Osburn, 1989). Validity arises from the extent to which scores on predictors and criteria share the same underlying constructs (Tenopyr, 1977).

Thorough explication of the constructs underlying scores on tests designed to sample specific KSAs required for job performance increases scientific understanding of tests and provides a basis for judgments about the tests' potential validity in new situations. Assessment of the constructs in measurement may also indicate that alternative tests provide essentially parallel information about applicants' standing on a given construct. This may provide the basis for the development of new parallel tests that can then be substituted for older measures of the constructs to maintain test security (Turban et al., 1989). To the extent that replacement measures of the same constructs are designed to reflect the KSAs required in the jobs at hand, the benefits described previously may be realized in addition to the reduction of possible problems with the security of existing tests.

Of interest in the present study was the extent of overlap in the constructs measured by commercially available aptitude tests and tests designed to measure the KSAs required in the performance of specific jobs. An approach to assessing the construct equivalence of test scores (Jöreskog & Sörbom, 1989) recently applied in a selection context by Turban et al. (1989) was adopted in this study. In this approach, confirmatory factor analysis is used to determine the extent to which new replacement tests measure the same underlying constructs with equivalent strength and uniquenesses as existing measures. To the extent that replacement tests are equivalent measures of the constructs underlying performance on existing measures, the new tests may serve as substitutes for the original tests in measuring the focal construct. New replacement tests may be written to reflect job-specific KSAs, or, alternatively, commercially available tests, which are economically more desirable and generalizable across different situations, may be adopted to avoid the expense of developing job-specific tests.

A thorough assessment of the construct equivalence of tests may also include an examination of the extent to which constructs are measured equivalently across various subgroups (Jöreskog & Sörbom, 1989; see also Drasgow & Kanfer, 1985; Turban et al., 1989). In the present study, equivalent constructs were hypothesized to underlie scores on tests taken from commercially available aptitude test batteries and tests designed to sample the specific KSAs required for performance in several jobs. The equivalence of these constructs across various gender and ethnic subgroups was also assessed. Several models, differing in their assumptions about factor loadings and variable uniquenesses, were tested to assess the construct equivalence of the job-specific and commercially available tests. An examination of the equivalence of constructs measured by the tests should increase scientific understanding of the tests and allow a determination of the extent to which the measures can be substituted for one another with no loss of information about the underlying construct.

Method

Sample

The study sample included 3,956 applicants for a journey-level apprenticeship training program at a large manufacturing firm. The sample comprised 2,691 White men, 694 African-American men, 291 White women, 172 African-American women, and 108 Hispanic men. Data were collected during the actual hiring process at locations throughout the midwestern, eastern, and southeastern United States. Applicants sought positions in apprenticeship training programs in one of eight skilled trades: tool and die maker, tool maker, die maker, machine repair, millwright, plumber/pipefitter, electrician, and welder. A portion of the sample (n = 306) was included in a criterion-related validation study reported elsewhere (Hattrup & Schmitt, 1990).

Measures

On the basis of analyses of the task and KSA requirements of each of the eight jobs, several paper-and-pencil tests commercially available as part of published test batteries were obtained for use in selecting applicants for training. Job analysis also provided the basis for the construction of additional paper-and-pencil tests that included items representing the KSAs required in each of the eight trades. The job analyses consisted of extensive interviews with job experts, followed by task surveys to determine the importance and frequency of performance of various tasks by apprentices in the eight trades. Group interviews were then conducted with experienced journey-level workers to identify the KSAs required on each task. These KSA–task linkages were used to select the commercially available tests and job-specific measures described in the following paragraphs. At least in terms of their development, the three job-specific tests are fairly typical of other such specially constructed tests used in selection contexts.

Though not designed to be parallel forms or measures of the constructs underlying the commercially available tests, several of the job-specific tests appeared to measure constructs similar to those measured by the aptitude tests. Descriptions of the constructs and the corresponding commercially available and job-specific tests follow.

Verbal reasoning. The Verbal Reasoning subscale of the Differential Aptitude Test (DAT) battery (Bennett, Seashore, & Wesman, 1972) measures the ability to understand and use verbally expressed relationships. The subscale is composed of 50 verbal analogies. The internal consistency (alpha) of this measure in the validation subsample (for which item-level data were available) was .91. A job-specific measure, labelled Technical Reading, was developed; it contained 25 reading comprehension items referring to passages describing job-relevant technical information. The internal consistency of this measure was .87.

Numerical ability. This construct represents the ability to understand and manipulate quantitative information. The Numerical Ability subscale of the DAT, which contains 40 items dealing with addition, subtraction, multiplication, division, percentages, and ratios, was used. Its reliability in the validation sample was .91. A job-specific

subtest labelled Industrial Math was created; it contained eight items dealing with common measurement and arithmetic problems encountered in these jobs. The internal consistency of this measure was .71.

Spatial visualization. This is a construct referring to the ability to mentally represent three-dimensional space and relationships among objects. The Space Visualization subtest of the Employee Aptitude Survey (EAS; Grimsky, Ruch, Warren, & Ford, 1957) contains 50 items that refer to relationships among blocks in a diagram. Its reliability was .94. The job-specific test was labeled Following Instructions and contains 20 items referring to the row and column location or the color (or both) of circles arranged in numbered row × column matrices. This test was meant to simulate the kinds of instructions appearing in technical books, manufacturers' manuals, and other materials used by the trade workers. The internal consistency of this test was .85.

Analytic Procedure

Figure 1 presents the hypothesized model relating the manifest variables and underlying constructs. The factor structure, and maximum likelihood estimates of factor loadings, factor intercorrelations, and unique residual variances were examined with LISREL VII (Jöreskog & Sörbom, 1989). LISREL provides the simultaneous assessment of relationships in a hypothesized model. In confirmatory factor analysis, LISREL allows the examination of models that are based on various assumptions about factor loadings, factor intercorrelations, and uniquenesses that are fixed to be equal to specific values or to other parameter estimates, or are allowed to be free to take on any value. Three models of the relationships between these six tests, each based on different notions of test parallelism, were assessed with LISREL.

The first, least restrictive model assumes only that the true scores of equivalent tests are perfectly correlated, though true scores may differ by an additive or multiplicative constant. This model is termed the congeneric (Linn & Werts, 1979), or essentially tau-equivalent (Lord & Novick, 1968), model of equivalence. Because the congeneric model does not require equal error variances, one of the tests may be measured with greater reliability. The factor analytic corollary of congenericism requires only that the patterns of factor loadings be equal for the two batteries of tests examined. In other words, the congeneric model implies that equivalent tests load on a shared construct, although the loadings may differ in magnitude. In the present study, the factor analytic equivalent of the congeneric model required that the patterns of factor loadings conform to the pattern illustrated in Figure 1.

The second model, termed tau equivalence, assumes that tests have equal true score variance but not necessarily equal error variance or reliability (Lord & Novick, 1968). Tau equivalence is a special case of congenericism that implies equal factor loadings of tests hypothesized to measure identical constructs. Tau-equivalent tests measure the same construct to the same degree but are not equally reliable.

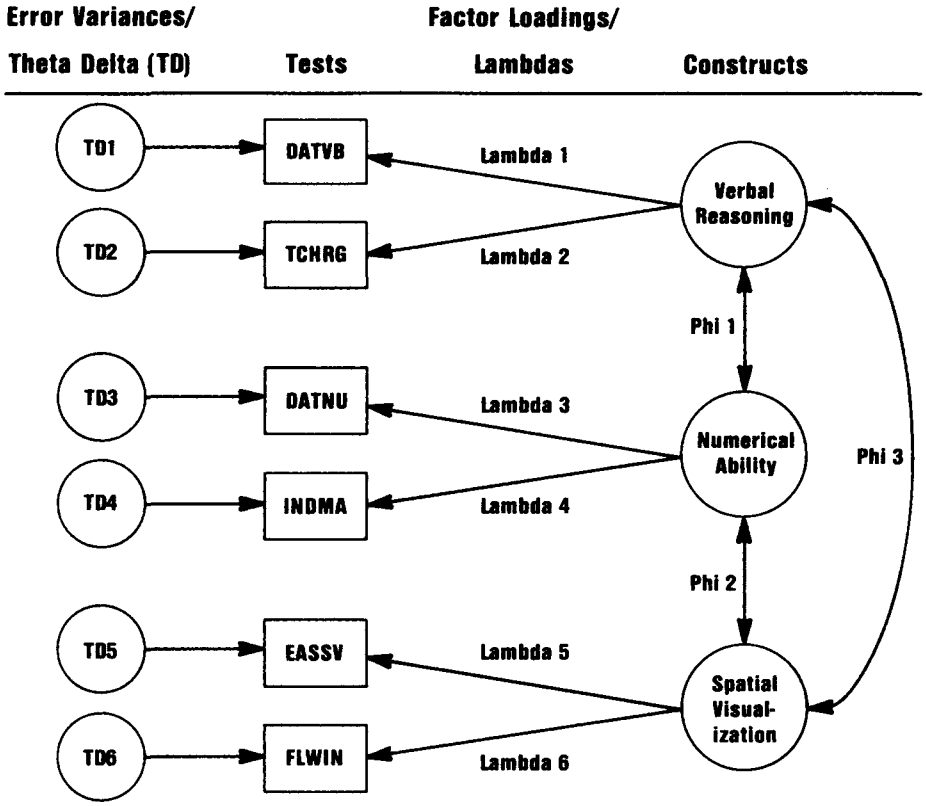
Finally, tests are said to be parallel if true score variances, error variances, and true scores of the two tests are equal (Lord & Novick, 1968). Hence, parallel tests load on the same construct to the same degree and have equal error. However, because error variances are independent (although of equal magnitude), parallel measures will not correlate perfectly with each other (Allen & Yen, 1979). Validities of parallel tests in predicting an external criterion should be equivalent, but validities of tests that are tau-equivalent or congeneric will not be equal because of unequal reliabilities. Observed scores are interchangeable only when the parallel model holds. Parallelism further requires that these restrictions hold in each subsample drawn from a given population of test takers (Lord & Novick, 1968). The factor analytic equivalent of parallelism implies equal factor loadings and equal unique variances among tests hypothesized to be equivalent.

The constraints implied by different versions of the factor analytic model of equivalence are identical to the constraints implied by classical test theory notions of congenericism, tau equivalence, and parallelism only when there is no unique but reliable variance in each test. That is, the congeneric model assumes that only nonsystematic sources of variability contribute to the unique component in each test, whereas the factor analytic approach assumes that uniquenesses may contain both systematic sources of variance that are specific to a given measure and nonsystematic errors of measurement (Linn & Werts, 1979). Hence, although parallel measures correlate equally with other tests, tests that satisfy the factor analytic equivalent of parallelism may not correlate equally with an external criterion if the measures contain specific variance. Strictly speaking, confirmatory factor analysis indicates only the extent to which tests measure equivalent constructs and not their parallelism in the classical test theory sense unless it can be demonstrated that tests contain no specific and systematic variance.[1]

The three models of equivalence are hierarchically nested, with the congeneric model being the least restrictive and the parallel model being most restrictive. Hence, if the factor analytic corollary of the parallel model provides an adequate fit to the data, the remaining models will also fit. Each model was tested hierarchically in this study in order from least to most restrictive. Evidence that a less restrictive model fit the data provided a basis for testing models that incorporated greater restrictions. The congeneric equivalent (H_{form}) was tested first by simply constraining the factor patterns to conform to the model illustrated in Figure 1. In other words, the commercially available and job-specific tests in a given pair were assumed to load on a single construct hypothesized to underlie scores on those measures. This was followed by a test of the tau-equivalent corollary (H_λ), which further required freely estimated but equal loadings of tests in a pair on the latent constructs represented in the lambda matrix. Finally, the parallel equivalent model ($H_{\lambda\Theta}$) was tested by further fixing the uniquenesses implied by the theta–delta matrix to be equal for the tests in each of the three pairs of commercially available and job-specific measures. Although tests incorporating added restrictions are unnecessary when less restricted models fail to fit the data, tests of all three models were performed to provide estimates of the goodness of fit or importance of constraints associated with the tau-equivalent and parallel models.

To assess the extent to which these models fit consistently across race and sex subgroups, we used LISREL VII's (Jöreskog & Sörbom, 1989) multiple-groups confirmatory factor analysis. Tests of model invariance across groups can vary in restrictiveness (cf. Bollen, 1989);

[1] One method of assessing the extent of correspondence between classical test theory assumptions of equivalence and the factor analytic corollary in a given data set involves comparing each test's alpha reliability to $1 - (\Theta_\delta/\sigma^2_x)$, where Θ_δ is the maximum likelihood estimate of the test's unique variance, and σ^2_x is the test's total variance. The difference between these values is an estimate of the proportion of total variance that is specific and systematic but unrelated to the underlying construct estimated with LISREL. On the basis of the theta values (uniquenesses) presented in Table 3, test-specific variance in the present study ranged from .57 for the Space Visualization measure to .10 for the Following Instructions test, with an average across the six tests of .24. Hence, although the confirmatory factor analytic approach used here does not support statements of test equivalence in the classical test theory sense, the labels congeneric, tau-equivalent, and parallel are retained in the text for brevity and to maintain consistency with previous research (Jöreskog & Sörbom, 1989; Turban et al., 1989).



NULL MODEL: Lambda 1 = Lambda 2 = Lambda 3 = Lambda 4 = Lambda 5 = Lambda 6 = Phi 1 = Phi 2 = Phi 3 = 0.

CONGENERIC MODEL: Relationship as indicated by figure. Lambdas and theta deltas not constrained to be equal.

TAU-EQUIVALENT MODEL: Lambda 1 = Lambda 2; Lamba 3 = Lambda 4; Lambda 5 = Lambda 6.

PARALLEL MODEL: Lambda 1 = Lambda 2; Lambda 3 = Lambda 4; Lambda 5 = Lambda 6; TD1 = TD2; TD3 = TD4; TD5 = TD6.

NOTE: DATVB = Differential Aptitude Test Verbal Reasoning, TCHRG = Job Relevant Technical Reading, DATNU = Differential Aptitude Test Numerical Ability, INDMA = Job Relevant Industrial Math, EASSV = Employee Aptitude Survey Space Visualization, FLWIN = Job Relevant Following Instructions Test.

Figure 1. Hypothesized model of the relationships among constructs underlying scores on job-specific and commercially available tests (adapted from Turban, Sanders, Francis, & Osburn, 1989).

however, five models were consistent with the implications of equivalence outlined previously. A sixth model added restrictions relevant to the invariance of relations among constructs. The first step in assessing model invariance across the five race and sex subgroups (i.e., White men, African-American men, White women, African-American women, and Hispanic men) involved a test of the assumption of equal covariances (H_{cov}) across the groups (see Jöreskog & Sörbom, 1989). This required computation of rescaled subgroup variance–covariance S_g matrices as described by Jöreskog (1971). Such rescaling involves dividing each element of the subgroup covariance matrices by the corresponding across-group, weighted-average standard deviations. Rescaling in this manner eliminates between-test differences in observed score variances but retains important differences between groups in the variances of observed scores and factor scores (see Cudeck, 1989).

If the assumption of equality of S_g matrices does not hold across the

groups, an assessment of the equivalence of the congeneric, tau-equivalent, and parallel models across groups then provides information about what components (factor patterns, factor loadings, or uniquenesses) of testing show differences across the groups (Drasgow & Kanfer, 1985; Turban et al., 1989). Although an assessment of less restrictive models is unnecessary if the model dictating equal rescaled covariances shows adequate fit, all tests were performed and are reported here.

The second step, therefore, in the multiple-groups analysis involved assessing the invariance of the congeneric equivalent (H_{form}), which assumed equivalent factor patterns, both within test pairs and across the five subgroups. A third test constrained equal factor loadings of tests within each pair (i.e., the tau-equivalent model, H_{λ}) and across the five subgroups. A fourth test involved further restrictions consistent with the factor analytic corollary of the parallel model ($H_{\lambda\Theta}$). These restrictions included equal factor patterns, factor loadings, and uniquenesses both within test pairs and across the five subgroups. Finally, a model constraining equal factor patterns, loadings, uniquenesses, and factor intercorrelations across the groups represented the most restricted possibility ($H_{\lambda\Theta\Phi}$). This final model did not incorporate additional restrictions relevant to test equivalence. However, it was included to test the hypothesis that correlations among underlying constructs were invariant across groups.

To assess the appropriateness of the various models, we examined several indices of fit. LISREL VII provides goodness-of-fit (GFI), adjusted goodness-of-fit (AGFI), and root-mean-square residual (rmsr) indices for most tests involving models and model invariance. In the case of multiple-groups confirmatory factor analysis, LISREL VII provides GFI and rmsr values for each subgroup in the analysis, indicating the invariance of the observed group covariance matrix relative to a reproduced matrix implied by estimated model parameters and based on a weighted combination of subgroup matrices. All three indices have been shown to be influenced by variations in sample size (Marsh, Balla, & McDonald, 1988). Therefore, using the GFI or rmsr values to evaluate the relative fit of models across groups is of limited usefulness. Nevertheless, all indices are presented here for comparison.

LISREL VII also provides an overall chi-square value indexing the extent to which the observed covariance or correlation matrix differs from the reproduced matrix implied by the estimated model parameters. In the case of multiple-groups confirmatory factor analysis, the chi-square represents the fit of the model in all groups taken simultaneously. However, because the chi-square value is dependent on sample size (Schmitt & Stults, 1986), large samples may result in significant chi-squares even for appropriate models. Therefore, the nonnormed fit index (NNFI) described by Tucker and Lewis (1973; see also Bentler & Bonett, 1980; Marsh et al., 1988) was also computed to assess the appropriateness of each model. The NNFI is derived by comparing the chi-square/degrees of freedom ratio of each model tested with that of a null model (H_{null}) specifying zero covariances among measures and sample invariant variances. By convention, NNFI values exceeding .90 indicate acceptable fit.

The NNFI has been shown to be relatively insensitive to variation in sample size (Marsh et al., 1988). Furthermore, although increasing the number of parameters to be estimated generally increases the value of other fit indices, such as the GFI provided by LISREL VII, the NNFI controls for the number of estimated parameters by including the degrees of freedom of models in its computation (Marsh et al., 1988). Hence, more parsimonious models (i.e., those with fewer estimated parameters) will yield higher NNFI values when all else is equal.

In the present study, rescaled covariance matrices were analyzed to assess the construct equivalence of the commercially available and job-specific measures. However, analyses of S_g matrices or of correlation matrices cannot formally be used to determine parallelism or tau equivalence (Cudeck, 1989; Turban et al., 1989). This is because parallelism and tau equivalence refer to the equivalence of tests in their original units of measurement, not standardized units. Parallelism refers to the extent to which observed scores on tests are interchangeable. Therefore, tests that are congeneric and have equal reliabilities will fit a parallel model when they are standardized (Kenny, 1979; Turban et al., 1989). Pairs of tests in this study differed greatly in their means and standard deviations. It was not expected therefore, that analyses of covariance matrices would provide meaningful information about the equivalence of tests that would be standardized before comparison. When the practical problem is one of test substitution or construct explication, and not of comparison of observed scores that are assumed to be in the same metric, the analysis of rescaled or standardized matrices is appropriate.

In this study, both multiple-groups confirmatory factor analyses and analyses of the fit of the congeneric, tau-equivalent, and parallel corollaries in the total sample were performed. In practice, assessment of the fit of models in a combined sample is appropriate only after demonstrating model invariance across subgroups. Nevertheless, results for the total sample are reported first for clarity of presentation. To control for the effects of subgroup mean differences on the observed total sample correlation matrix, we pooled subgroup S_g matrices before analyzing the fit of models in the combined sample. The weighted average of subgroup S_g matrices results in a pooled correlation matrix for the total sample (Jöreskog, 1971). The subgroup observed and rescaled variance–covariance matrices are available on request.

Results

Observed Correlations

Descriptive statistics, including means, standard deviations, and the unpooled and pooled total sample intercorrelations of the six tests are presented in Table 1. Table 1 also presents the disattenuated (corrected for unreliability in the tests) correlations based on the observed unpooled correlation matrix. The pattern of disattenuated correlations provided some initial evidence of the construct equivalence of the commercially available and job-specific measures. As can be seen in the table, several of the correlations involving tests hypothesized to measure the same construct were higher than correlations among tests hypothesized to measure distinct constructs. For example, the largest disattenuated correlation involving the Numerical Ability test was with the Industrial Math test, and the largest disattenuated correlation involving the Technical Reading test was with the Verbal Reasoning measure. Similar relationships can be noted among the observed and pooled correlations. The most notable exception was the Following Instructions test, which was less highly correlated with the Space Visualization test than with any of the other tests. One possible explanation is that scores on the Following Instructions test may be highly influenced by general cognitive ability, more so than on the Space Visualization test. This would result in large and uniform correlations of the Following Instructions test with any other measure that is also strongly influenced by general ability. Although Space Visualization is probably also influenced by general cognitive ability, disattenuated correlations involving this test were the lowest obtained, suggesting that this test was

Table 1
Means, Standard Deviations, and Unpooled and Pooled Intercorrelations of Commercially Available and Job-Specific Tests for the Total Sample

Test	*M*	*SD*	1	2	3	4	5	6
			Unpooled correlations[a]					
1. Verbal Reasoning	21.25	8.58	—	.78	.61	.82	.77	.84
2. Numerical Ability	16.71	9.28	.71	—	.62	.71	.87	.78
3. Space Visualization	24.23	11.03	.56	.57	—	.59	.67	.72
4. Technical Reading	19.59	6.74	.73	.63	.53	—	.76	.81
5. Industrial Math	5.72	2.20	.62	.70	.55	.60	—	.88
6. Following Instructions	14.71	5.80	.74	.69	.64	.70	.68	—
			Pooled S_g matrix					
1. Verbal Reasoning			—					
2. Numerical Ability			.66	—				
3. Space Visualization			.47	.48	—			
4. Technical Reading			.68	.57	.42	—		
5. Industrial Math			.54	.64	.44	.51	—	
6. Following Instructions			.69	.63	.53	.63	.60	—

Note. $N = 3{,}956$.
[a] Values below the diagonal are observed correlations; values above the diagonal are unpooled correlations corrected for unreliability in the tests.

somewhat less related to a general factor underlying scores on all of the tests.

Confirmatory Factor Analysis

The fit indices, including the GFI, AGFI, rmsr, and NNFI, for the null model and the factor analytic corollaries of the congeneric, tau-equivalent, and parallel models for the total pooled sample are reported in the top half of Table 2. As can be seen in the table, the fit indices for models based on the pooled S_g matrix indicated that all three models of equivalence provided a good fit to the observed data. AGFI and NNFI values exceeded .90, and rmsr values were low for each of the models tested. Chi-square values were large for each of the models; however, interpretation of these values is problematic because of their dependence on sample size. Although the results supported the fit of all three equivalence models, statistically significant differences in chi-square values, and practical differences in NNFI values, suggest that the congeneric model provides a superior fit to the data.

Table 3 presents the LISREL maximum likelihood estimated parameters for the least restricted equivalence model (i.e., the congeneric model) for each subgroup and for the pooled sample. The adequate NNFI values for each of the equivalence models based on the pooled matrix support the conclusion that tests in each of the pairs of commercially available and job-specific tests measure equivalent constructs with equal uniquenesses. However, as can be seen in Table 3, large differences in the estimated uniquenesses of the Space Visualization and Following Instructions tests suggest that a model allowing freely estimated theta-deltas for these two tests but equal uniquenesses among the other two pairs would provide a better fit to the data relative to the fully constrained parallel model. Fit indices of this alternative model (Model 1), presented in the bottom half of Table 2, indicate superior fit relative to the fully constrained parallel model (NNFI = .949 vs. .925 for the parallel model, chi-square difference = 285.26, $p < .001$). These results were not entirely surprising, given the lower observed correlations involving the Space Visualization test.

Substantial observed correlations among the six tests, and large estimated intercorrelations among the underlying factors, suggested that an alternative model that constrained equal loadings on a single common factor and equivalent uniquenesses might fit the data. Such a model would be consistent with a single underlying general factor on which all tests load uniformly. As can be seen in Table 2, this alternative model (Model 2) fit the data nearly as well as the fully constrained parallel model. However, the chi-square difference between the

Table 2
Confirmatory Factor Analysis of Construct Equivalence of Commerically Available and Job-Specific Tests

Model	χ^2	*df*	NNFI	GFI	AGFI	rmsr
H_{null}	12,319.18	15	—	.379	.131	.484
Equivalence model						
H_{form} (congeneric)	104.08	6	.980	.991	.970	.014
H_{Λ} (tau equivalent)	383.59	9	.949	.970	.929	.088
$H_{\Lambda\Theta}$ (parallel)	752.18	12	.925	.945	.903	.064
Alternative model						
Model 1	466.82	11	.949	.964	.932	.084
Model 2	1,123.81	19	.929	.912	.902	.073

Note. Analysis is based on the pooled S_g matrix. NNFI = nonnormed fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; and rmsr = root-mean-square residual. Model 1 refers to a model that was identical to the parallel model except that separate uniquenesses were estimated for the Space Visualization and Following Instructions tests. Model 2 was a model in which equal factor loadings on a single common construct and equivalent uniquenesses were specified.

one-factor and three-factor parallel models was significant (chi-square difference = 371.63, $p < .001$), indicating that part of the substantial fit of the one-factor model, at least on the basis of its NNFI, was due to its greater parsimony. Additional single-factor models with relaxed assumptions for equality constraints provided less adequate fit than did their analogous three-factor equivalence models.

Combined with an examination of the observed intercorrelations among the six tests, the results of confirmatory factor analysis indicate that the least restrictive model of equivalence with three underlying factors provided the closest fit to the observed data. Tests in each pair loaded on their shared construct as illustrated in Figure 1. Large NNFI values for the tau-equivalent corollary and for the alternative model specify-

Table 3
Maximum Likelihood Estimates of Model Parameters Associated With the Congeneric Equivalent

	White men			African-American men			White women			African-American women			Hispanic men			Total sample		
Test	VR	NA	SV	VR	NA	SV	VR	NA	SV	VR	NA	SV	VR	NA	SV	VR	NA	SV
								Factor loadings										
Verbal Reasoning	0.88			0.81			1.02			0.71			0.88			0.87		
Numerical Ability		0.90			0.72			0.84			0.69			0.87			0.85	
Space Visualization			0.59			0.59			0.83			0.60			0.74			0.61
Technical Reading	0.74			0.88			0.87			0.66			1.09			0.78		
Industrial Math		0.67			1.00			0.85			0.86			1.03			0.76	
Following Instructions			0.82			0.99			0.99			0.82			1.15			0.87
								Factor intercorrelations										
Factor																		
Verbal Reasoning	—			—			—			—			—			—		
Numerical Ability	.85	—		.88	—		.87	—		.86	—		.88	—		.87	—	
Space Visualization	.91	.89	—	.92	.88	—	.89	.88	—	.94	.85	—	.94	.83	—	.92	.89	—
								Variable uniquenesses										
Test																		
Verbal Reasoning		.25			.24			0.26			.18			.19			.25	
Numerical Reasoning		.29			.19			.33			.16			.24			.28	
Space Visualization		.63			.60			0.59			.51			.76			.63	
Technical Reading		.36			.45			0.37			.56			.31			.39	
Industrial Math		.38			.50			0.45			.41			.55			.43	
Following Instructions		.24			.25			.23			.29			.17			.25	

Note. VR = Verbal Reasoning factor; NA = Numerical Ability factor; Space Visualization factor.

ing freely estimated uniquenesses for the space visualization tests but equal uniquenesses among tests in the other two pairs also indicated acceptable practical fit of these models. Hence, tests in each pair can be considered to assess their common constructs with equivalent strength and, in the case of the verbal and numerical ability test pairs, with equivalent uniquenesses. Observed scores are expected to differ substantially, as the test lengths, means, and variances are not uniform within each pair. Moreover, reliabilities of the tests also may differ, particularly among the two numerical ability tests, as evidenced by their alpha coefficients and the disparity between estimated alphas and LISREL-estimated common variances (see Footnote 1).

Multiple-Groups Analyses

The fit indices for the tests of invariance of the null and alternative models, based on subgroup S_g matrices, are presented in Table 4. As noted previously, LISREL VII provides an overall chi-square indicating the fit of the model in all groups taken simultaneously. Nonnormed fit indices were also computed for each overall model and are displayed in Table 4. Also presented in Table 4 are GFI and rmsr indices for each subgroup in the analysis, indicating the invariance of each subgroup's rescaled covariance matrix relative to the combined matrices across groups. Because the subgroup GFI and rmsr values may be influenced by variation in sample size or in rescaled variances, the NNFI values for each overall model provide the most justifiable standard against which to compare models.

As can be seen in Table 4, NNFI values indicate that the congeneric and tau-equivalent models and the model constraining equivalent covariances across the groups fit adequately when subgroup S_g matrices were analyzed. The model constraining equal covariances (H_{cov}) provided the best fit to the data (NNFI = .959), in part because of its greater parsimony. The factor analytic corollaries of the congeneric and tau-equivalent models also resulted in NNFI values exceeding the .90 cutoff set by convention. Although the NNFI for the fully constrained model that restricted equal loadings, uniquenesses, and factor correlations exceeded .90, the similarity of its subgroup GFI and rmsr values and its overall chi-square value to values for the parallel model suggest that greater fit was due to parsimony in the number of estimated parameters.

Chi-square differences were significant between all of the models tested except the two most restrictive possibilities. Hence, given practical differences in the size of each model's NNFI, the congeneric equivalent provided the closest fit to the data. The tau-equivalent model provided adequate but somewhat poorer fit relative to the congeneric equivalent. GFI values for the five sex and race subgroups were above .90 for all of the groups tested for the congeneric model, and were above .90 for all groups except the Hispanic male group for the tau-equivalent model. As can be seen in Table 3, estimated factor loadings in the Hispanic male group for the Technical Reading and Following Instructions tests were substantially larger than corresponding loadings in the other four groups. However, the GFI values were somewhat lower for the Hispanic male subgroup than for the other groups in all of the models tested. These

Table 4
Multiple-Groups Confirmatory Factor Analysis of Construct Equivalence

Model	χ^2	*df*	NNFI	White men		African-American men		White women		African-American women		Hispanic men	
				GFI	rmsr	GFI	rmsr	GFI	rmsr	GFI	rmsr	GFI	rmsr
H_{null}	12,785.74	99	—	.386	.456	.353	.561	.331	.635	.391	.421	.300	.748
Equivalence model													
H_{cov}	521.63	84	.959	.985	.070	.909	.180	.955	.191	.895	.155	.852	.331
H_{form}	117.77	30	.977	.991	.014	.998	.009	.984	.019	.981	.027	.964	.034
H_{Λ}	664.05	57	.917	.953	.112	.933	.157	.951	.190	.953	.139	.876	.305
$H_{\Lambda\Theta}$	1,266.24	84	.890	.926	.099	.863	.190	.929	.203	.847	.161	.813	.341
$H_{\Lambda\Theta\Phi}$	1,274.01	96	.904	.926	.098	.863	.190	.928	.195	.845	.158	.810	.337
Alternative model													
Model 1	968.94	83	.917	.952	.118	.866	.202	.944	.188	.854	.182	.832	.329
Model 2	1,629.77	103	.884	.896	.104	.844	.192	.900	.197	.839	.157	.783	.339

Note. Analyses are based on the subgroup S_g matrices. NNFI = nonnormed fit index; GFI = goodness-of-fit index; rmsr = root-mean-square residual; H_{null} = null model; H_{form} = congeneric model; H_{Λ} = tau-equivalent model; $H_{\Lambda\Theta}$ = parallel model; and $H_{\Lambda\Theta\Phi}$ = a model including invariant factor loadings, factor intercorrelations, and variable uniquenesses. Model 1 refers to a model that was identical to the parallel model except that separate uniquenesses were estimated for the Space Visualization and Following Instructions tests. Model 2 was a model in which equal factor loadings on a single common construct were specified.

lower values may be due in part to the smaller sample size of the Hispanic male group relative to the other subgroups (cf. Marsh et al., 1988).

A lower NNFI for the factor analytic corollary of the parallel model indicated poor fit of this model across the groups. Subgroup GFI values were relatively low for the African- American male, African-American female, and Hispanic male groups. To assess whether the poor fit was due to differences in the uniquenesses of the two space visualization tests, we analyzed an alternative model that allowed freely estimated but invariant thetas for the these two tests and equivalent and invariant uniquenesses among tests in the other two pairs. As can be seen in Table 4, this alternative model (Model 1) improved the fit of the fully constrained model, at least in terms of the overall NNFI value. However, subgroup GFI values remained low for the racial minority groups. Most notably, a low GFI was obtained for the African-American male group despite a larger sample relative to the White female group.

Subgroup GFI values for Model 1 approached the values obtained from the tau-equivalent corollary for the White male and female groups but did not do so for the racial minority groups. Hence, a model that constrained equivalent and invariant factor loadings but freely estimated uniqueness between tests and across the groups was the most restrictive model that also fit the data. A similar picture can be seen by examining the variable uniquenesses in Table 3. Estimated thetas for the White male and White female groups were similar to each other and to the results of the pooled S_g analysis, whereas thetas for the minority groups differed for several of the tests. Differences in the minority groups' uniquenesses may suggest differences in the reliabilities or specific variances in these groups.[2]

Because of the uniformly large estimated factor intercorrelations in each of the five subgroups, a single-factor model with equivalent loadings and uniquenesses of all six tests and invariance in these parameters across the groups was also assessed. As displayed in Table 4, this single-factor model (Model 2) provided a poor fit to the observed data. Relaxing equality constraints among tests in the lambda and theta matrices and invariance of subgroup loadings in the least restrictive model resulted in poorer fit than analogous models based on three factors. Hence, LISREL multiple-groups confirmatory factor analysis support the conclusion that tests loaded on their hypothesized shared constructs, as illustrated in Figure 1, and that the three constructs can be considered distinct. Constraining equivalent factor loadings within test pairs and invariance of loadings across the groups also provided adequate practical fit, though not superior to the less restrictive congeneric model. Again, observed scores on tests would be expected to differ, and tests may contain differential specific variance or reliabilities across the five subgroups analyzed in this study.

Discussion

In this study, we explored the equivalence of constructs measured by selection tests designed to reflect job-specific KSAs with the constructs underlying scores on tests taken directly from published test batteries. The results indicated that, when standardized, the job-specific Technical Reading, Industrial Math, and Following Instructions tests provided tau-equivalent measures of the primary constructs assessed with the published Verbal Reasoning, Numerical Ability, and Space Visualization tests. With the exception of the Space Visualization and Following Instructions measures, tests were also composed of equivalent uniquenesses that were unrelated to the underlying common constructs.

Large differences in the means and standard deviations of tests meant that observed scores on the job-specific tests and commercially available measures could not fit a model constraining equivalent factor loadings or uniquenesses. In other words, the observed scores on the two types of tests cannot be substituted for each other without first standardizing the scores on each test. Because the practical problem addressed in this study was the equivalence of constructs, and not the substitutability of observed scores, results using rescaled and standardized scores suggested strong support for our expectations about the constructs hypothesized to underlie scores on the different tests. In practical terms, this implies that the job-specific tests could replace the commercially available tests with little loss of information about the focal constructs.

Substantial uniquenesses in each test, and large discrepancies in the size of these uniquenesses relative to the estimated alpha reliabilities, however, suspend any strong conclusions about the equivalence of tests' validities in predicting an external criterion. Validities may differ by a value that is directly related to the specific proportion of variance in a test, assuming that all of the variance in the external criterion is related to the specific variance, an unlikely possibility. Thus, the results support conclusions about the equivalence with which underlying constructs are measured by the tests. Because some of the tests (Space Visualization, in particular) contained large portions of specific but reliable variance, there may be some difference in the validities of these tests. Actual observed validities of the tests with job sample criteria in a portion of the present sample (Hattrup & Schmitt, 1990) varied somewhat by test, though not in a manner that could be interpreted apart from sampling error alone.

Multiple-groups analyses also indicated that the tests assessed their underlying constructs equivalently across various subgroups. When rescaled covariances for the subgroups were analyzed, several restrictive models, which required sample invariant covariances among tests (H_{cov}) and invariant factor loadings (H_λ), provided a very good fit to the observed rescaled matrices. Although there was some evidence that a model that further constrained equivalent and sample invariant uniquenesses among the verbal and numerical tests, but freely estimated uniquenesses among the two visualization tests, resulted in adequate overall fit, subgroup GFI values indicated that uniquenesses were quite varied across the groups. This was especially true among the racial minority groups (i.e. African-American men, African-American women, and Hispanic men). Hence, although tests loaded equivalently on shared constructs across the groups, remaining unique variance in the tests appeared to vary across the groups, particularly among

[2] Unfortunately, the unavailability of subgroup samples containing full item data precluded the computation of separate subgroup alpha reliabilities.

the racial minorities. This would indicate differing reliabilities or specific variances (or both) across subgroups and, to the extent that specific variances are differentially related to an external criterion of job performance, possible bias across the groups. However, differential prediction analyses using a portion of the present sample and based on composites formed from the job-specific and commercially available test batteries resulted in nonsignificant slope differences in the prediction of job sample criteria (Hattrup & Schmitt, 1990). Further research of the constructs underlying tests, including the ways in which constructs are measured equivalently across groups, and whether subgroup-specific variances are differentially related to important performance-related variables, appears warranted.

The results of this study point to the possibility that tests designed to measure specific KSAs required in the performance of given jobs may, in fact, assess the same constructs represented in more general aptitude tests. As described previously, such job-specific measures possess many advantages in terms of user acceptance and perceived fairness. Although tests designed to sample job-specific requirements may be costly to develop, they are generally viewed more favorably by test takers and the courts (Kleiman & Faley, 1985; Schmitt & Klimoski, 1991). Hence, the benefits accrued from such favorable impressions may outweigh the early costs in test development.

Of course, such differences in user acceptance of the present tests, if they do in fact exist, can be attributable only to perceived differences between the tests and not to differences in the actual adverse impact of the tests. Mean differences between groups were relatively constant and large (about one standard deviation between African-Americans and Whites, and half of a standard deviation between men and women) on each of the commercially available and job-specific measures. Unfortunately, no data were available on perceptions of test fairness in the present context. Nevertheless, negative reactions to tests that lack face validity have been documented (e.g., Schmidt et al., 1977) and can be expected in many testing situations. Research is needed that assesses both the reactions to job-specific and commercially available tests and the constructs and other sources of variance that underlie scores on these measures.

On the other hand, the results reported in this article support the conclusion that one may gain little psychometrically by developing job-specific exams. Of course, we report the results of only one set of comparisons between general and specific tests. Replication of our results with other tests should be undertaken. As one reviewer of our article noted, there is a whole universe of constructs that might be examined in this way. Although the three constructs we examined may be fairly representative of the cognitive-ability domain, the job-specific tests developed in this research might have been quite different had other psychologists analyzed the jobs and written the test items. An obvious difference might be the use of test formats other than paper-and-pencil, as is common with many job-specific exams. Outside the cognitive-ability realm, there may be no similar convergence among personality inventories and job-specific assessment centers or situational interviews. Given other constructs, tests, and situations, there is no assurance that job-specific and generic tests would match as well as they did in this study.

A related issue has to do with the large intercorrelations among all of the tests examined in this research. Given this circumstance, nearly any randomly chosen pair of tests might exhibit some degree of parallelism. However, an evaluation of the degree to which a single common factor accounted for all covariation among tests (see Tables 2 and 4) indicated significantly poorer fit of single-factor models relative to models in which tests loaded on their hypothesized distinct factors. One would expect larger differences between a single-factor model and multifactor models if constructs were less highly interrelated, as was the case in this study.

The present study, and similar research reported by Turban et al. (1989), suggests new avenues for research into the construct validity of measurement devices. The construct equivalence technique described here (see also Drasgow & Kanfer, 1985; Turban et al., 1989) may provide a useful approach to exploring the constructs underlying scores on various other selection and attitude measures. The results of such investigations may serve to broaden the scientific understanding of test scores and the extent to which tests measure constructs equivalently across groups or contain subgroup-specific variance.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1972). *Differential Aptitude Tests.* New York: Psychological Corporation.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analyses of covariance structures. *Psychological Bulletin, 88,* 588–606.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

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

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26,* 461–477.

Grimsky, G. L., Ruch, F. L., Warren, N. D., & Ford, J. S. (1957). *Employee Aptitude Survey: Space Visualization.* Los Angeles, CA: Psychological Services.

Guion, R. M. (1978). "Content validity" in moderation. *Personnel Psychology, 31,* 205–213.

Hattrup, K., & Schmitt, N. (1990). Prediction of trades apprentices' performance on job sample criteria. *Personnel Psychology, 43,* 453–466.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL VII: A guide to the program and applications* (2nd ed.). Chicago: SPSS, Inc.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Kleiman, L. S., & Faley, R. H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. *Personnel Psychology, 38,* 803–833.

Linn, P. C., & Werts, C. E. (1979). Covariance structures and their analysis. In R. Traub (Ed.), *New directions for testing and measurement: Methodological developments* (pp. 53–73). San Francisco, CA: Jossey-Bass.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187–198.

Schmitt, N., & Klimoski, R. J. (1991). *Research methods in human resources management.* Cincinnati, OH: Southwestern.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait–multimethod matrices. *Applied Psychological Measurement, 10,* 1–22.

Smither, J. W., & Pearlman, K. (1991, April). Perceptions of the job-relatedness of selection procedures among college recruits and recruiting/employment managers. In R. R. Reilly (Chair), *Perceived validity of selection procedures: Implications for organizations.* Symposium conducted at the Sixth Annual Conference of the Society for Industrial and Organizational Psychology, St.Louis, MO.

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

Tenopyr, M. L. (1977). Content–construct confusion. *Personnel Psychology, 30,* 47–54.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Turban, D. B., Sanders, P. A., Francis, D. J., & Osburn, H. G. (1989). Construct equivalence as an approach to replacing validated cognitive ability selection tests. *Journal of Applied Psychology, 74,* 62–71.

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

Received February 28, 1991
Revision received November 15, 1991
Accepted November 22, 1991 ■