Journal of Applied Psychology
1989, Vol. 74, No. 1, 62-71

0021-9010/89/$00.75

Construct Equivalence as an Approach to Replacing Validated Cognitive Ability Selection Tests

Daniel B. Turban
University of Houston

Patricia A. Sanders
Shell Oil Company

David J. Francis and H. G. Osburn
University of Houston

In this study we demonstrate an approach to replacing validated selection tests to which job applicants may have prior access. This approach, labeled construct equivalence, allows for replacing valid tests currently in use with new, experimental tests that have been shown to measure the same constructs. We demonstrated the construct equivalence approach by collecting data from over 2,000 applicants for four different positions in a large petrochemical company. We investigated the equivalence of the experimental and the current tests by using correlational analyses, structural modeling, and analyses of hiring decisions. Results indicated that the experimental and current tests measure the same constructs and that replacing the current tests with the experimental tests would treat ethnic and sex subgroups consistently. Construct equivalence was shown to be a viable approach to test substitution.

Maintaining the security of selection tests is a practical problem for both publishers and test users (Guion, 1987). Although most test publishers sell tests only to qualified users, publishers have virtually no control over the purchaser's actual use (or misuse) of test materials (Holmen & Docter, 1972). Test security may be further jeopardized by the widespread use of certain tests within industries. For example, many petrochemical companies with employment offices in the same cities use, or have used, a common set of cognitive ability tests for personnel selection (Callender & Osburn, 1981; Schmidt, Hunter, & Caplan, 1981). In addition, many organizations have used the same set of tests for personnel selection for an extended period of time. Consequently, many organizations are faced with the problem that some applicants may have prior information concerning these selection tests and some may even have access to the correct answers to test items. This is a concern not only because of the unfair advantage for applicants with prior access but also because the validity of the tests may be substantially reduced.

Unfortunately, when it becomes necessary to replace current, validated cognitive ability tests with new, experimental tests, the options are often limited. One possible solution is to conduct criterion-related validity studies of new, secure tests. However, concurrent validity studies with incumbents are often not feasible owing to severe restriction of range resulting from high minimum scores on the current cognitive test batteries. Predictive validity studies, which can minimize restriction of range, typically are not practical owing to low hiring rates and the negative consequences of hiring low-scoring applicants.

Of course, if secure, parallel alternate forms were available, these could be substituted for the current forms without compromising the validity of selection procedures. Unfortunately, in many situations, secure, parallel forms do not exist. In that case, another possible solution is for the organization to develop alternative measures of the current, validated tests. However, many organizations lack the resources necessary to develop their own tests and must therefore rely on published tests. Selecting items from an item bank and administering different, but equivalent, tests to applicants is another solution that poses similar difficulties, because few organizations maintain item banks or have the resources necessary to analyze item-level data.

An approach that many organizations should find feasible is to replace the current, validated tests with secure, experimental tests that measure the same constructs or combination of constructs. This approach, labeled construct equivalence, is based on the idea that (a) it is the ability, or trait, measured by the current test that is related to job performance, in other words, the test's predictive validity comes from its relation to some underlying construct (ability or trait) and the role of that construct in job performance; and that (b) new tests can be found that measure the same constructs as the current tests, even though these tests were not constructed as strictly parallel alternate forms of the current tests. The construct equivalence approach is derived from construct validity theory (Cronbach & Meehl, 1955) and involves administering concurrently both current and experimental tests to applicants to determine whether the new, experimental tests measure the same constructs as the current, validated tests. If both the current and new tests measure the same construct, or combination of constructs, then the experimental test can replace the previously validated test in the

Portions of this article were presented at the 95th Annual Convention of the American Psychological Association, August 28–September 1, 1987, New York. We gratefully acknowledge the helpful comments of Allan Jones, Edward Kahn, Angela McDermott, and Carol Timmreck on earlier drafts of the manuscript.

Correspondence concerning this article should be addressed to Daniel B. Turban, Psychology Department, University of Houston, Houston, Texas 77004.

selection battery. Whether a construct is conceptualized as a singular construct or a combination of constructs does not affect the construct equivalence approach to test substitution, because if two tests measure the same construct, or combination of constructs, then the new test can replace the current test in the selection battery. Therefore, for simplicity in presentation, the term construct is used to refer to singular constructs and a combination of constructs.

A construct is conceptualized as a "postulated attribute assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). In the present study, we assume that variance in current test scores is caused by individual differences in the construct measured by the test and that another test may also measure that same construct. If it is assumed that the construct reflected in the test score influences job performance, then another measure of that construct would also presumably predict job performance. Therefore, evidence for the validity of the new tests in predicting job performance is obtained by demonstrating that a new test measures the same construct as the current test.

In the present study, we deal with the practical problem of replacing current, validated tests by demonstrating an approach to investigating whether new, experimental tests measure the same constructs as the current tests and can therefore replace the current tests in a selection battery.

Method

Sample

The sample consisted of 2,430 applicants for clerical, secretarial, exploration and production (E&P) operating, and refinery and chemical plant (R&C) process operating positions in a large petrochemical company. Data were collected at five different sites: three in the Southeast, and one each in the Southwest and Midwest. Data for the clerical and secretarial positions were collected from two sites, data for the R&C positions were collected from two other sites, and data for the E&P positions were collected from the fifth site. We collected these data during an actual hiring process, and the sex and ethnicity of applicants were representative of previous applicants for these positions.

Procedure

Applicants were tested as part of an ongoing hiring process and therefore had reason to believe that all test results would be used for selection purposes. The current tests were administered first, followed by the experimental tests. The testing sessions lasted approximately 2–4 hr, depending on the job group.

Constructs

The three constructs measured by the previously validated, current tests and the experimental tests were general mental ability, mechanical reasoning, and spatial visualization. For secretarial and clerical positions, the construct was general mental ability; for E&P positions, the constructs were general mental ability and mechanical reasoning; and for R&C operating positions, the constructs were general mental ability, mechanical reasoning, and spatial visualization.

General mental ability is the ability to manipulate abstract verbal, numerical, and figural symbols. This construct is currently measured by a 12-min test, referred to as GMA, containing 18 block-counting problems, 18 vocabulary items, and 18 arithmetic problems in a spiral omnibus pattern. The experimental test, designated EGMA, is a 40-min test containing 25 verbal comprehension items, 25 verbal reasoning items, 15 figural reasoning items, and 15 quantitative reasoning items also in a spiral omnibus pattern.

Table 1
Descriptive Statistics and Intercorrelations of Current and Experimental Tests for Total Sample

Test	*N*	GMA	MR	SV	EGMA	EMR	ESV	*M*	*SD*
GMA	2,430	—						28.40	8.18
MR	1,509	.553	—					40.33	10.44
SV	1,244	.520	.675	—				27.25	11.05
EGMA	2,429	.788	.656	.640	—			36.37	15.54
EMR	1,507	.541	.868	.676	.658	—		45.71	12.55
ESV	1,242	.528	.707	.901	.656	.719	—	32.48	12.43

Note. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively.

Mechanical reasoning is the ability to perceive and understand the relation of physical forces and mechanical elements in practical situations. The current test for this construct, identified as MR, contains 68 multiple-choice items that realistically illustrate situations involving mechanical devices or principles. The experimental test, EMR, contains 70 items that are very similar to items included in the MR. Both tests have the same 30-min time limit.

Spatial visualization is the ability to visualize a three-dimensional object from a picture of the pattern of its surfaces and to imagine how the object would appear if rotated in various ways. The current test for this construct, referred to as SV, and the experimental test, ESV, each contain 60 of the same type of items and have the same 25-min time limit.

As far as we can determine, none of the experimental tests were specifically designed to be parallel, alternate forms of the current tests and have not been validated for the job groups investigated. The experimental and the current tests of spatial visualization and mechanical reasoning are very similar to each other and were published by the same company, however, different companies published the experimental and the current general mental ability tests. Table 1 presents descriptive statistics and intercorrelations of the current and experimental tests for the total sample.

Analyses and Results

Correlations

One way to investigate whether the current and experimental tests measure the same construct is to correct both tests for unreliability to estimate the correlation between true scores on the tests. A corrected correlation of 1.00 suggests that the tests are equivalent measures of the construct. Table 2 presents uncorrected and corrected correlations of the paired current and experimental tests by job, sex, and ethnic group. The correlations were corrected for unreliability in both tests (Thorndike, 1982). The total sample corrected correlations between the current and experimental tests of the general mental ability, mechanical reasoning, and spatial visualization constructs were .877, 1.00, and .953, respectively. The corrected correlations analyzed by job, sex, and ethnic subgroups indicate similar results. Overall, these results suggest that MR and EMR measure one construct;

Table 2
Correlations of Paired Current and Experimental Tests for Job, Sex, and Ethnic Subgroups

Group	GMA with EGMA			MR with EMR			SV with ESV		
	n	*r*	r^c	*n*	*r*	r^c	*n*	*r*	r^c
Job									
Clerical	415	.749	.834						
Secretarial	504	.747	.831						
E&P operations	265	.792	.881	265	.841	.972			
R&C operations	1,241	.814	.906	1,241	.870	1.000	1,241	.901	.953
Sex									
Female	1,119	.759	.845	282	.754	.872	279	.868	.919
Male	1,306	.808	.899	1,224	.857	.991	962	.904	.957
Ethnic									
White	1,204	.757	.842	841	.839	.970	646	.903	.956
Black	1,090	.746	.830	643	.774	.895	574	.852	.902
Total minority	1,221	.743	.827	665	.781	.903	595	.851	.901
Total sample	2,425	.788	.877	1,506	.868	1.00	1,241	.901	.953

Note. E&P = exploration and production; R&C = refinery and chemical. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. r^cs are correlations corrected for unreliability in both measures. Reliability coefficients used for the corrections were: GMA = .85, EGMA = .95; MR = .86, EMR = .87; SV = .95, and ESV = .94.

SV and ESV measure one construct; and GMA and EGMA measure very similar but possibly not identical constructs.

Structural Analysis

The corrected correlations provide initial evidence concerning the equivalence of the current and the experimental tests, however, this method has limitations. Internal consistency reliability estimates were obtained from published reports. Although the largest applicable reliability estimates were used in the corrections, these estimates may over- or underestimate reliability in a given sample. The accuracy of reliability estimates is a concern because underestimated reliability coefficients artificially inflate the corrected correlations. Furthermore, as typically applied, this method focuses only on relation between tests and therefore loses information about relations among all tests and constructs. Although one could disattenuate all correlations and subsequently examine them for perfect within-construct correlations and low to moderate cross-construct correlations, there is no easy method for testing that such a pattern holds on the disattenuated correlations. To overcome these limitations and to investigate the psychometric properties of all tests simultaneously, we used structural modeling analyses. Structural modeling is a particularly powerful technique to investigate whether tests measure the same constructs because (a) the reliability estimates are based on the sample under investigation and the particular model specified; (b) the user can specify the relation between constructs and tests; (c) the relations among all tests and constructs can be investigated simultaneously, therefore using more information than the correlational analyses; (d) the degree of equivalence of tests can vary across constructs; and (e) the equivalence of tests can be examined simultaneously in multiple groups. These strengths allow investigation into many different models of the equivalence of paired current and experimental tests.[1]

We conducted the structural analyses with LISREL VI (Jöreskog & Sörbom, 1986). LISREL VI is a general statistical program that can be used for confirmatory factor analysis and allows for comparison of different models of the relations between the tests and constructs by allowing a user to (a) fix parameters to an assigned value, (b) constrain parameters to be equal to an unknown value of one or more other parameters, and (c) allow free parameters that are not constrained to be equal to any other parameter or value (Jöreskog & Sörbom, 1986). LISREL VI provides users with the ability to compare statistically the goodness of fit of different models of relations among constructs and tests with the obtained data. Only data from the refinery and chemical plant process operations applicants were analyzed with LISREL VI for two major reasons: (a) It is the largest job group, and (b) it is the only job group that was administered all three experimental tests.[2]

[1] To use structural modeling to investigate the equivalence of tests, the models must be overidentified. Identification is a complex issue beyond the scope of this article and is generally a function of the number of tests, the number of constructs, and whether the constructs are correlated. Interested readers are referred to Kenny (1979).

[2] The data used in the structural analyses were the uncorrected covariances and are presented in the Appendix. The analyses in this section, based on combined ethnic groups (i.e., the total R&C sample), assume that models are comparable across ethnic groups. The approach demonstrated in the next section tests that assumption. In practice, researchers should consider the subgroup analyses before the total group analyses. However, in many situations, sufficiently large samples will not be available in all subgroups to make the subgroup approach feasible. For that reason, we demonstrated both approaches in the present study. Furthermore, when the sample consists of subgroups that are known to have mean differences on the tests (e.g., Blacks and Whites), structural analyses on the total sample should be conducted on a pooled within-groups covariance matrix so as to remove the effects of mean differences on estimated covariances. Although analyses of the pooled

Psychometric theory distinguishes three levels of test equivalence (see Figure 1). Congeneric tests may have unequal true score variances and unequal error variances; however, the relation between congeneric measures "is attributable to a single underlying latent variable" that "accounts for all of the systematic variability of the observed measures" (Linn & Werts, 1979, p. 55). Thus, in this context, congenericism implies a three-factor model in which the relation between each pair of current and experimental tests is caused by only one construct. At the very least, the data must fit the congeneric model for construct equivalence to be demonstrated. Tau-equivalent and parallel tests are considered special cases of congeneric tests that place additional constraints on error variances and true score variances (Jöreskog, 1979a). Tau-equivalent tests have equal true score variances, but unequal error variances, whereas parallel tests have equal true score variances and equal error variances. Thus, as indicated in Figure 1, the tau-equivalent model constrained the factor loadings to be equal for each pair of tests, and the parallel tests model constrained both the factor loadings and the error variances to be equal for each pair of current and experimental tests. Finally, the null model is used only for comparative purposes and states that there is no relation among the tests, in other words, the variance in each test is error variance.

We investigated three psychometric models because these models have different implications for test substitution. Observed scores on congeneric tests reflect the same construct, although not necessarily to the same degree, and not necessarily as reliably. Tau-equivalent tests measure a construct to the same degree, yet are not equally reliable. Therefore, true scores on tau-equivalent tests differ by a constant and may be measured more accurately by one of the tau-equivalent tests than by the other (Allen & Yen, 1979). Finally, parallel tests measure the same construct to the same degree and are equally reliable. Therefore, parallel tests have equal observed score means and equal correlations with other variables. Only when tests are parallel, will the validity of the new tests equal the validity of the current tests. If the tests are congeneric or tau-equivalent, the new tests may be more or less valid for predicting job performance than the current tests, in part because reliabilities may differ. In addition, scores on parallel tests are interchangeable. However, observed scores on congeneric or tau-equivalent tests are not interchangeable and would therefore necessitate rescaling the raw scores before making selection decisions.

Because the three substantive models are hierarchically nested from most restrictive (parallel) to least restrictive (congeneric), more than one of these models may fit the data. Specifically, if a restrictive model fits the data, then all less restrictive models must also fit. Consequently, researchers should not necessarily stop the analyses when the results indicate that a specific model fits the data. Rather, all three substantive models should be assessed to determine the effect of each set of restrictions. Because the restrictions are nested, specified in advance, and represent theoretically defensible positions, determining which model fits the data is considered a confirmatory, rather than an exploratory, analysis (Loehlin, 1987).

Table 3 presents results from structural analyses of both covariances and correlations. Strictly speaking, tests cannot be determined to be parallel or tau-equivalent by analyzing correlations, because parallelism and tau-equivalence are statements about tests in their original metric, not in a standardized metric. However, the tests in this study would not be expected to be strictly parallel because of their different lengths, although they may be tau-equivalent. More importantly, if the tests are congeneric and have equal reliabilities, then the tests will be parallel when standardized (Kenny, 1979). Furthermore, because in this study we addressed the practical problem of test substitution and were not concerned with whether the true score was in the units of measurement of the tests (Kenny, 1979), we analyzed all models with both covariances and correlations.

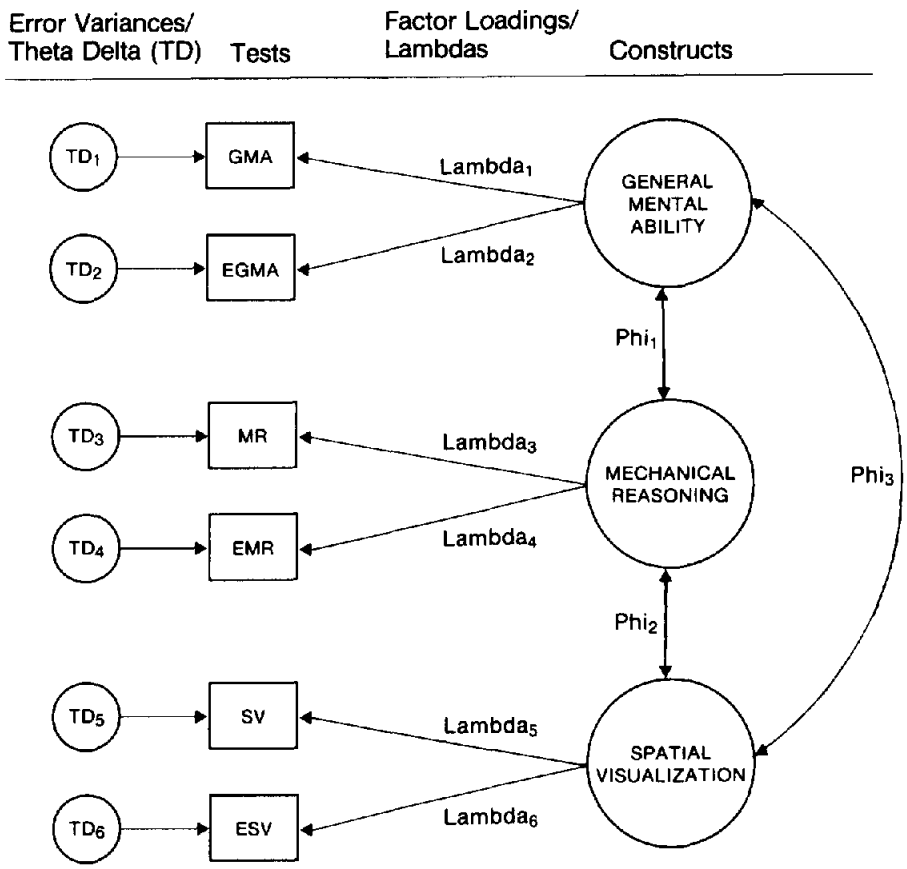
The results indicate that the congeneric tests model is most suitable for these data. Although the chi-square is statistically significant ($p \leq .008$), this is not surprising given the large sample size and, as Jöreskog (1979b) notes, "a model may well be accepted even though χ^2 is large" (p. 201). The large size of the goodness-of-fit and normed fit indices, and the small size of the root mean square residual, indicate that the congeneric tests model provides a very good fit to the data. When the covariance matrix is analyzed, the parallel and tau-equivalent models do not fit the data; however, when the correlations are analyzed, these models provide an excellent fit to the data.

The pattern of corrected correlations presented in Table 2 suggests that perhaps the current and experimental tests of the general mental ability construct were congeneric, whereas the tests of the mechanical reasoning and spatial visualization constructs were either tau-equivalent or parallel. Therefore, we investigated these additional models (see Table 3). Although the congeneric model for all constructs necessarily fits the data best, the results suggest that for practical purposes the paired tests of the general mental ability construct are congeneric (i.e., one factor accounts for all the systematic variance in both tests), and the paired tests of the mechanical reasoning and spatial visualization constructs can be considered parallel (i.e., equal true scores and reliabilities). The correlational and structural modeling analyses provide strong evidence to indicate that, although the strength of the relation varies across constructs, each pair of experimental and current tests measures the same construct.

Ethnic and Sex Differences

Although considerable evidence indicates that the use of cognitive tests to predict job performance criteria is not unfair to minorities (Hunter, Schmidt, & Hunter, 1979; Schmidt, Pearlman, & Hunter, 1980), LISREL was used to determine whether the relation of the experimental and the current tests with the constructs is consistent across sex and ethnic subgroups. If the relation of the current and the experimental tests with the constructs are relatively consistent across groups, it indicates that the tests affect the groups similarly, at least within the context

within-groups covariance matrix slightly improved the fit of the models to the data in the current study, conclusions concerning the fit of the models were the same as when the unpooled covariance matrix was analyzed. Therefore, for ease of presentation, we presented the results on the total sample from analyses of unpooled covariance matrixes, with results from the total R&C sample presented first, followed by the subgroup analyses.



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: $Lambda_1 = Lambda_2$; $Lambda_3 = Lambda_4$; $Lambda_5 = Lambda_6$.

Parallel Model: $Lambda_1 = Lambda_2$; $Lambda_3 = Lambda_4$; $Lambda_5 = Lambda_6$; $TD_1 = TD_2$; $TD_3 = TD_4$; $TD_5 = TD_6$.

Figure 1. Structural models of the relation of the current and experimental tests with the latent constructs.

of these models. Therefore, we analyzed the R&C data with LISREL by investigating the fit of the data to a model that constrains certain parameters to be equal in both groups. However, different parameters can be constrained to be equal across groups, therefore we investigated different models.

The null model is used for comparative purposes only and constrains the variance of each test to be equal across groups, although there are no covariances among the tests in either group. The hypothesis of equal covariance matrixes constrains the variances and covariances to be equal in both groups. If this hypothesis is rejected, then it remains possible that the correlation matrixes (i.e., the standardized covariance matrixes) are equal in both groups. If neither covariance nor correlation matrixes are equal across groups, then the important question is, how are the groups unequal? The least restricted possibility that allows for measurement comparability across groups is the hypothesis of invariant factor patterns, in other words, the congeneric model holds in both groups, with the paired current and experimental tests measuring the same construct irrespective of group membership. If this hypothesis is tenable, then it may be appropriate to test more restricted hypotheses. The more restricted hypotheses test (a) whether the factor loadings are invariant across subgroups; (b) whether the factor loadings and error variances are invariant across subgroups; and (c) whether the factor loadings, error variances, and factor correlations are invariant across subgroups. If any of the less restricted hypothe-

Table 3
Confirmatory Factor Analysis of Relation Between Current and Experimental Tests

Model	χ^2	*df*	*p*	Goodness-of-fit index	Adjusted goodness-of-fit index	Root mean square residual	Normed fit index
			Covariance matrix				
Null	7,087.56	15	.0001	.299	−1.454[a]	83.746	—
Parallel tests	1,356.33	12	.0001	.774	.472	36.126	.809
Tau-equivalent tests	777.37	9	.0001	.834	.709	37.212	.890
Congeneric tests	17.48	6	.008	.995	.993	.867	.998
GMA congeneric and SV and MR tau-equivalent	189.87	8	.0001	.950	.919	10.192	.973
GMA congeneric and SV and MR parallel	275.96	10	.0001	.930	.866	11.412	.961
			Correlation matrix				
Null	7,087.56	15	.0001	.299	−1.454[a]	.579	—
Parallel tests	129.29	12	.0001	.967	.924	.035	.982
Tau-equivalent tests	73.15	9	.0001	.981	.966	.049	.990
Congeneric tests	17.48	6	.008	.995	.993	.008	.998
GMA congeneric and SV and MR tau-equivalent	24.56	8	.002	.993	.989	.014	.997
GMA congeneric and SV and MR parallel	35.75	10	.0001	.991	.983	.011	.995

Note. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. N = 1,241 for all analyses.
[a] Although the adjusted goodness-of-fit index is usually between zero and one, it is theoretically possible for it to become negative (Jöreskog & Sörbom, 1986). This indicates, as expected, that the null model does not fit the data.

ses are rejected, then the more restricted hypotheses will either be rejected or will be of questionable scientific use; for example, testing for equality of factor loadings is only meaningful when the tests measure the same factors in both groups. Nevertheless, we present the results from all analyses for interested readers (see Table 4).

These results indicate that the relation of the current and the experimental tests with the constructs are relatively consistent across ethnic and sex subgroups. For both the ethnic and sex subgroup analyses, the model with the best fit to the data states that the congeneric model fits in both groups. Although the chi-square is significant (at the $p \leq .05$ level) for the sex subgroups, the magnitude of the normed fit index indicates that this model fits the data extremely well. In addition, the factor loadings invariant model also fits the data extremely well for both the sex and ethnic subgroups. These results indicate that the current and the experimental tests measure the constructs similarly for ethnic and sex subgroups.

Operational Effects

The results indicate that the new and the current tests measure the same constructs and that the relationships of the current and the experimental tests with the constructs are similar across sex and ethnic subgroups. If the current and the new tests were strictly parallel, it would be unnecessary to investigate the operational effects of the substitution. However, because the results are equivocal about whether the tests are strictly parallel, we conducted the operational analyses to provide additional information about the construct equivalence of the tests. To conduct the operational analyses, we created battery scores, following the organization's normal procedures, by standardizing each test in the battery and summing the standardized scores. The clerical battery consisted of the GMA (EGMA) test and two other tests, the secretarial battery consisted of the GMA (EGMA) test and one other test, the E&P battery consisted of the GMA (EGMA) and the MR (EMR) tests, and the R&C battery consisted of the GMA (EGMA), MR (EMR), and SV (ESV) tests. The uncorrected correlations between the new and the old test battery scores for the clerical, secretarial, E&P, and R&C job groups were .96, .93, .88, and .94, respectively.

Further evidence concerning the operational effects of the substitution is provided by directly comparing decisions made under the two sets of tests. Figure 2 presents fourfold tables of pass and fail rates on the current and the experimental test batteries, analyzed by sex and ethnic subgroups. Although we expected that some individuals would pass one test battery and fail the other, we expected these mismatches to differ only as a result of chance sampling (McNemar, 1969). The nonsignificant chi-squares indicated there were no differences in the mis-

Table 4
Testing Hypotheses of the Equality of the Relations of the Current and Experimental Tests with the Constructs for Sex and Ethnic Subgroups

	Ethnic[a]			Sex[b]		
Hypothesis	χ^2	p	Normed fit index	χ^2	p	Normed fit index
Equal null model ($df = 36$)	5,752.11	.001	—	6,912.98	.001	—
Equal covariance matrixes ($df = 21$)	75.54	.001	.987	67.42	.001	.990
Equal correlation matrixes ($df = 21$)	82.80	.001	.986	74.21	.001	.989
Factor patterns invariant ($df = 12$)	16.07	.188	.997	22.61	.031	.997
Factor loadings invariant ($df = 15$)	32.01	.006	.994	44.26	.001	.994
Factor loadings and error variances invariant ($df = 21$)	66.26	.001	.988	60.79	.001	.991
Factor loadings, error variances, and factor covariances invariant ($df = 27$)	88.76	.001	.985	85.77	.001	.988

Note. All analyses were performed with covariances except for the equality of correlation matrixes hypothesis.
[a] Equality of the relationships of the current and experimental tests with the constructs for Black ($n = 574$) and White ($n = 646$) subgroups. [b]Equality of the relationships of the current and experimental tests with the constructs for female ($n = 279$) and male ($n = 962$) subgroups.

matches, that is, the number of applicants who passed the current but failed the experimental battery was equivalent to the number who passed the experimental but failed the current battery. These analyses also indicated that an equivalent proportion of applicants passed the current and the experimental test batteries within ethnic and sex subgroups.

The only test in the clerical and secretarial batteries with an experimental form was the measure of the general mental ability construct. Although previous analyses suggested that EGMA and GMA may measure the same construct slightly differently, the operational analyses (not shown in Figure 2) indicated that the number of mismatches did not differ significantly for clerical or secretarial applicants as a function of the test used. This indicates that the slight difference in measuring the construct does not affect resulting decisions.

For the total sample, less than 5% passed the current battery but failed the experimental battery, and over two-thirds of these applicants were within 3 scale points of the experimental battery cutoff score. Because similar results would be expected with a retest of the current battery, this provides further evidence that the experimental tests are equivalent measures of the constructs measured by the current tests and can therefore replace the current tests in the selection battery.

Discussion

In this study we demonstrated an approach to determine whether experimental tests can replace current, validated cognitive ability tests in selecting job applicants. This approach, labeled construct equivalence, states that if another test measures the same construct as the validated, current test, then this new test can replace the current test in the selection battery. We demonstrated the construct equivalence approach by administering current and experimental tests to job applicants. By using correlational, structural modeling, and operational effects analyses, we investigated the equivalence of the experimental and the current tests.

We used three types of analyses to investigate the equivalence of the tests because each analysis provides somewhat different information. Furthermore, the extent to which the three types of analyses lead to the same conclusions provides information about whether to replace the current test with the experimental tests. In the present study, because all three analyses led to the same conclusion, we decided that the experimental tests could replace the current tests in the selection battery.

The correlational analyses were based on the assumption that if the experimental and the current tests measure the same con-

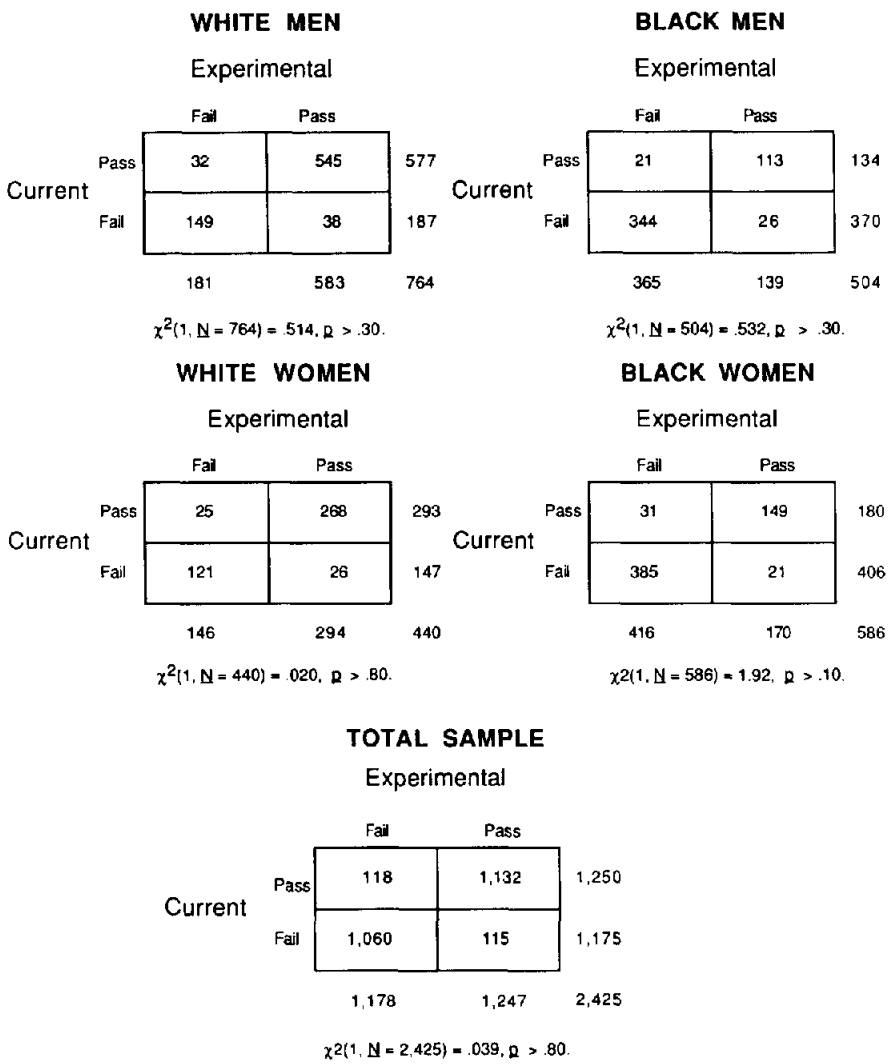


Figure 2. Pass and fail rates on the current and experimental test batteries across job groups: total sample and by sex and ethnic subgroups.

struct, then the true correlation between these tests would be 1.00, although any estimated true correlation may be different from 1.00 because of sampling error. Correlations of the paired experimental and current tests corrected for unreliability in both measures provided preliminary evidence that the experimental tests measure the same constructs as the current tests, although the strength of the relation varies across constructs.

Whereas the correlational analyses provided information about the relation between the paired tests, structural modeling (LISREL VI) investigated the relation of latent constructs with the tests. Although we analyzed both covariances and correlations, we deemed analyses of correlations most appropriate because standardized test scores are used for selection decisions, and this study deals with the practical issues of test substitution. The congeneric tests model fits the data best, although the model with congeneric tests of the general mental ability construct and parallel tests of the mechanical reasoning and spatial visualization constructs also provided an excellent fit to the data; the root mean square residual of .011 indicates that very little data were unexplained by this model. Because the experimental test of the general mental ability construct had a higher reliability estimate than the current test, it is expected to have greater validity in predicting job performance.

If the current and the experimental tests were strictly parallel, it would be unnecessary to investigate whether decisions made about applicants are consistent within subgroups. Operational analyses become increasingly important when insufficient subjects are available to allow subgroup analyses. In the present study, subgroup analyses using structural modeling indicated that the relations of the current and the experimental tests with the constructs were similar across sex and ethnic subgroups but that these measures, although essentially equivalent, may not be strictly parallel. Operational analyses indicated that the percentage of applicants passing the current and the experimental tests did not differ within sex and ethnic subgroups. Taken in sum, these results indicate that the experimental and the current tests treat ethnic and sex subgroups similarly.

Because the correlational and structural analyses indicated that the paired experimental and current tests measure essentially the same constructs, we expected that decisions made about applicants would be the same regardless of the test battery used. However, we conducted these analyses to provide further information about the effects of substituting the experimental tests. Results indicated that an equivalent proportion of applicants passed the current and the experimental test batteries; this provides further support for the substitutability of the experimental tests.

It is instructive to consider the outcome of the operational analyses, presented in Figure 2, when the current tests have been compromised. To the extent that a large number of applicants had prior access to the current tests, more applicants would be expected to pass the current battery and fail the experimental battery than would pass the experimental and fail the current battery. Thus, it is clear that construct equivalence studies should not be conducted at locations where many applicants have prior access to the current tests. Instead, data should be collected at locations where it is believed that the tests are secure and then, if results support construct equivalence, the experimental tests can be transported to other locations. Obviously, if many applicants have prior access to the test, results indicating a lack of construct equivalence may be obtained because the experimental and the current tests actually do not measure the same constructs; or because the current tests no longer measure the relevant constructs in the applicant pool in which many applicants have prior access to the current tests. We conducted the current study at locations believed to be secure; the results supported that belief.

The construct equivalence approach is not a panacea for breaches in test security. As noted previously, if many applicants have prior access to the current tests, it may not be feasible to conduct a study investigating the construct equivalence of tests. A construct equivalence study needs to be conducted before the tests are badly compromised or at locations where the tests are relatively secure; otherwise, results are uninterpretable. Furthermore, a construct equivalence study is not feasible if one cannot find new tests that, although not necessarily parallel, are relatively similar to the current tests. At the least, the new tests need to have item content similar to the current tests before one can empirically investigate the equivalence of the current and the experimental tests. Of course, it is not enough to state that two tests purport to measure the same construct; data are needed to determine whether the tests actually measure the same construct. In addition, a construct equivalence study can only be done after the current tests have been validated. Finally, if the job has substantially changed since the current tests were validated, one should not replace current tests (that may no longer be valid) with new tests of the same construct.

In summary, the construct equivalence approach suggests that once a criterion-related validation study indicates which constructs are related to job success, if the job remains essentially the same, alternate tests of the same construct may be substituted for current tests without the necessity of conducting new criterion-related validity studies. In this study we demonstrated that construct equivalence is a viable approach for investigating whether experimental tests can replace current tests in the selection battery. Advantages of the construct equivalence approach are as follows: (a) Data are collected from applicants during regular testing sessions; (b) the data required are only test scores, not item information; and (c) the statistical packages to analyze the data are readily available (e.g., SPSS-X, LISREL) in many organizations. Therefore, when alternate solutions are not feasible, organizations may wish to consider a construct equivalence approach to test substitution.

References

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

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Guion, R. M. (1987). Changing views for personnel selection research. *Personnel Psychology, 40,* 199–213.

Holmen, M. G., & Docter, R. (1972). *Educational and psychological testing: A study of the industry and its practices.* New York: Russell Sage.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721–735.

Jöreskog, K. G. (1979a). Analyzing psychological data by structural analysis of covariance matrices. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 45–100). Cambridge, MA: Abt Books.

Jöreskog, K. G. (1979b). Simultaneous factor analysis in several populations. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 189–206). Cambridge, MA: Abt Books.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods* (4th ed.). Mooresville, IN: Scientific Software.

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

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: Jossey-Bass.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the petroleum industry. *Journal of Applied Psychology, 66,* 261–273.

Schmidt, F. L., Pearlman, K. & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33,* 705–724.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Appendix

Covariances Used in the Structural Analyses

Table A1
Covariance Matrix for Total R&C Job Group

Test	GMA	MR	SV	EGMA	EMR	ESV
GMA	71.4543	—				
MR	50.9329	112.195	—			
SV	48.4748	78.9707	122.057	—		
EGMA	108.54	110.04	111.529	249.06	—	
EMR	59.3943	116.74	94.6553	130.884	160.548	—
ESV	55.5005	93.1335	123.861	128.754	113.365	154.724

Note. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. N = 1,241 for the total sample R&C job group.

Table A2
Covariance Matrix for Blacks

Test	GMA	MR	SV	EGMA	EMR	ESV
GMA	60.675	—				
MR	29.4184	70.1573	—			
SV	30.6018	43.6013	87.6463	—		
EGMA	76.261	55.1205	64.7521	165.639	—	
EMR	35.5879	69.2847	54.6886	69.5802	112.753	—
ESV	35.1716	50.7212	84.927	74.6926	65.6026	113.382

Note. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. N = 574 for the Black subgroup.

Table A3
Covariance Matrix for Whites

Test	GMA	MR	SV	EGMA	EMR	ESV
GMA	56.9735	—				
MR	31.8603	91.1369	—			
SV	34.2873	63.4671	115.849	—		
EGMA	88.7335	82.8507	92.7099	226.556	—	
EMR	36.5387	91.5946	76.3304	98.5813	125.561	—
ESV	38.1412	76.1537	115.463	106.522	92.8923	141.001

Note. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. N = 646 for the White subgroup.

Table A4
Covariance Matrix for Women

Test	GMA	MR	SV	EGMA	EMR	ESV
GMA	63.0543	—				
MR	33.0618	66.1629	—			
SV	38.1823	44.1707	86.4267	—		
EGMA	87.5596	73.0126	83.1033	201.89	—	
EMR	47.1732	66.5403	67.0332	104.141	117.909	—
ESV	47.44	56.92	90.2128	104.049	86.9316	124.844

Note. GMA = general mental ability; MR = mechanical reasoning; SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. N = 279 for the female subgroup.

Table A5
Covariance Matrix for Men

Test	GMA	MR	SV	EGMA	EMR	ESV
GMA	73.0231	—				
MR	51.3069	100.498	—			
SV	49.2085	77.23	126.865	—		
EGMA	112.562	109.669	114.57	257.975	—	
EMR	57.1531	101.126	88.428	125.275	136.617	—
ESV	55.0684	89.0832	126.805	129.519	103.522	155.019

Note. GMA = general mental ability, MR = mechanical reasoning, SV = spatial visualization. EGMA, EMR, and ESV = experimental GMA, MR, and SV, respectively. N = 962 for the male subgroup.

Received October 27, 1987
Revision received May 11, 1988
Accepted April 20, 1988 ■