Journal of Applied Psychology
1990, Vol. 75, No. 2, 175-184

0021-9010/90/$00.75

Biographical Data in Employment Selection: Can Validities Be Made Generalizable?

Hannah R. Rothstein
Department of Management
Baruch College, City University of New York

Frank L. Schmidt
Department of Management and Organizations
University of Iowa

Frank W. Erwin
Richardson, Bellows, Henry & Company
Washington, DC

William A. Owens
Institute for Behavioral Research
University of Georgia

C. Paul Sparks
Serendipity Unlimited
Houston, TX

The hypothesis was examined that organizational specificity of biodata validity results from the methods typically used to select and key items. In this study, items were initially screened for job relevance, keying was based on large samples from multiple organizations, and items were retained only if they showed validity across organizations. Cross-validation was performed on approximately 11,000 first-line supervisors in 79 organizations. The resulting validities were meta-analyzed across organizations, age levels, sex, and levels of education, supervisory experience, and company tenure. In all cases, validities were generalizable. Validities were also stable across time and did not appear to stem from measurement of knowledge, skills, or abilities acquired through job experience. Finally, these results provide additional evidence against the hypothesis of situational specificity of validities, the first large-sample evidence in a noncognitive domain.

Substantial evidence now indicates that the two most valid predictors of job performance are cognitive ability tests and biodata instruments. The quantitative review of the literature by Hunter and Hunter (1984) has estimated the average validity of tests of general cognitive ability against supervisory ratings of overall job performance as .47, whereas the average (cross-validated) biodata validity against the same criterion was estimated as .37. Other review authors have obtained similar estimates of average (cross-validated) biodata validity. Reilly and Chao (1982) found a mean validity of .35, and Asher (1972) reported that 90% of biodata validities in his review were above .30. Thus, both general cognitive ability and biodata instruments have substantial validity; however, are the validities generalizable? Research over the last decade has demonstrated that the validities of cognitive ability tests can be generalized across settings, organizations, and even different jobs (e.g., Dunnette et al., 1982; Hunter, 1980; Lilienthal & Pearlman, 1983; Pearlman, Schmidt, & Hunter, 1980; Schmitt, Gooding, Noe, & Kirsh, 1984; see also Hartigan & Wigdor, 1989). However, it is widely believed that the validities of empirically keyed biographical data scales are situationally specific. For example, Hunter and Hunter (1984) stated, "there is evidence to suggest that biodata keys are not transportable" (p. 89). Thayer (1977) argued that biodata validity is moderated by age, organizational practices and procedures, the criterion used, temporal changes in the nature of the job, and other factors. Dreher and Sackett (1983) have commented that despite sizable validities, the fact of organizational (and subgroup) specific items and keys precludes the possibility of generalized validities in the case of biodata. In our experience, many have expressed the belief that organizationally specific validities are inevitable in the case of biodata.

It is clear that the pioneers of biodata research believed that appropriately developed biodata forms would show generalizability (Campbell, Dunnette, Lawler, & Weick, 1970; Owens, 1968, 1976; Sparks, 1983). A strong emphasis of the well-known Standard Oil of New Jersey's (SONJ) Early Identification of Management Potential (EIMP) study was on the common core of all management activities rather than on narrow functional specialties. The biodata instrument that resulted from this study had validities that were generalizable across varied functions within SONJ and five affiliate companies (Campbell et al., 1970). A later study (Laurent, 1970) showed that the validities generalized to different (non-English-speaking) countries as well. Similarly, the Aptitude Index Battery (AIB), used in the insurance industry, was developed by a central research group (LIMRA) for use in many different life insurance companies (Thayer, 1977) and showed validity across many insurance companies (although Brown, 1981, found evidence that validity is somewhat higher in "better managed" companies).

The most complete theory of biodata validity, that proposed by Owens (Owens, 1968, 1976; Owens & Schoenfeldt, 1979),

We thank John Haymaker and Cathy Choisser for their assistance in making the data available, and John Hunter for useful suggestions on an earlier draft. Any remaining errors are those of the authors.

Correspondence concerning this article should be addressed to Hannah R. Rothstein, Department of Management, Box 507, Baruch College, 17 Lexington Avenue, New York, New York 10010.

also emphasizes the potential generalizability of the method. Owens's assessment–classification model assigns persons to membership in relatively homogeneous life-history subgroups. Group membership is determined by one's pattern of scores on 13 biodata factors. Membership in these groups has been found to be differentially related to performance and satisfaction in various kinds of work. Thus, general life experience factors are related (differentially) to performance in a variety of different jobs.

We hypothesized that organizational specificity of biodata validities is traceable to the methods used to select and key items for the final scale (i.e., the method of scale construction) rather than to any inherent inability of biodata scores to yield generalizable validities. Items are typically selected and keyed on the basis of samples from a single organization; as a result, items whose validity does not generalize across organizations may not be detected and eliminated.

In this article we describe biodata research in which a different approach was taken to constructing and keying biodata items. Item selection and keying were based on samples from multiple organizations; only items that performed adequately across organizations were retained in the final scale. Cross-validation of the final key was performed on a sample of approximately 11,000 first-line supervisors working in 79 different organizations. The resulting validities were subjected to meta-analysis to determine the generalizability of the validities of the biodata scale. The central hypothesis addressed in this research is the proposition that biodata validities are intrinsically specific to organizations; therefore, the critical meta-analyses are those across organizations. But the generalizability of the biodata validities across other potential moderators that have been hypothesized (age, race, sex, education, experience, and tenure) is also examined. Although we are planning research to do so, this study does not examine the factor structure or dimensionality of the biodata scale, nor does it focus on determining the precise psychological meaning of scores on the biodata scale.

Method

Biodata Instrument

The instrument investigated was the empirically keyed autobiographical component of the Supervisory Profile Record (SPR). The Supervisory Profile Record is described in detail in Richardson, Bellows, Henry & Co., Inc. (1981). The complete SPR consists of a judgment questionnaire in addition to the biodata questionnaire. The prototype SPR, which contained 99 judgment and 128 autobiographical items, was modeled after instruments shown to be successful predictors of performance in the EIMP study. The judgment items were developed to obtain each respondent's views on such important content areas as employee motivation, personnel training and development, people and production problem resolution, discipline, and general supervisory style and practice. The autobiographical items were designed to elicit information about early developmental influences, academic history and accomplishments, and work-related values and attitudes (see next paragraph). Item development also included steps to ensure low reading difficulty levels and nonsexist language. Finally, each item was based on a rationale or psychological hypothesis, and all items were required to be such that prior supervisory experience was not required to respond to the item. This study focuses only on the biodata subscale.

The biodata key used in this study scores items from two rationally clustered subscales: (a) present self-concept/evaluation and (b) present work-values/orientation. For an individual to score high on the present self-concept evaluation scale he or she would (a) have a pervasive feeling of self-worth and confidence; (b) believe that he or she works better and faster than others in his or her area of specialization; (c) be recognized for accomplishments; (d) be outgoing, a good communicator, and a person who takes clear positions; and (e) feel healthy and satisfied with current life situations. The following are two items from that scale:

The amount of recognition which I usually receive for my accomplishments is:
(a) none at all
(b) occasional recognition, but not much
(c) about as much as deserved
(d) as much as deserved
(e) sometimes more than deserved

Of the following statements, the one which describes me best is:
(a) much more talker than listener
(b) somewhat more talker than listener
(c) about as much talker as listener
(d) somewhat more listener than talker
(e) much more listener than talker

An individual scoring high on present work-values/orientation is one who prefers to work independently, and who values hard work, drive, organization of time, and work planning. This person (a) works well under pressure, (b) is comfortable working on more than one thing at a time, and (c) views himself or herself as making progress and expects to continue to do so. The following are two items from that scale:

The kind of supervision I like best is:
(a) very close supervision
(b) fairly close supervision
(c) moderate supervision
(d) minimal supervision
(e) no supervision

My work habits are such that I prefer:
(a) to work on one thing at a time
(b) to work on several things at a time
(c) to work on many things at a time

The scale score used in our research is the sum of all keyed item scores.

Developmental Samples

Scale development. Item and key development was based on data from five consortia (SPR I, SPR II, SPR III, SPR IV, and SPR V), collectively labeled Consortium Grouping A in this study, with a total sample size of about 10,000 from 39 organizations. Each consortium consists of a group of organizations that participated in research on the SPR at a particular time. The first operational version of the SPR, developed after extensive review by psychologists, managers, and subject matter experts, was administered to a sample of 2,010 first-line supervisors from a consortium (SPR I) representing 336 locations of 6 organizations in a variety of industries. On the basis of results obtained with this sample, a tentative key was developed (see "Scoring Key Development," described later), revisions were made, and a second operational version was administered to a sample of 3,017 supervisors in an additional 10 organizations (SPR II). The third sample of 2,806 supervisors from 8 organizations made up SPR III. SPR V, with 659 supervisors, was the first completely white-collar sample and, despite the numbering, was begun and completed before SPR IV. All consortia were concurrent in nature. SPR IV, with 2,476 supervisors, represented the final developmental sample. No data from the first three consortia (SPR I–III), Consortium V, or most of SPR IV were included in this validity generalization study. The last 940 participants in Consortium IV were not used in the development of the biodata key studied in this research and were

Table 1
Developmental Sample Characteristics: Race, Sex and Collar (Consortium Grouping A)

	SPR I		SPR II		SPR III		SPR IV		SPR V		Total	
Composition	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Whites	1,702	85.3	2,311	92.7	2,064	91.5	2,369	95.5	368	82.5	8,814	91.2
Blacks	224	11.2	129	5.2	145	6.4	99	4.0	62	13.9	659	6.8
Hispanics	32	1.6	23	0.9	22	1.0	7	0.3	1	0.2	85	0.9
Other ethnic	38	1.9	24	1.0	25	1.0	6	0.2	15	3.3	108	1.1
Men	1,995	99.7	2,454	98.4	2,202	97.3	2,357	95.0	78	17.5	9,086	93.8
Women	6	0.3	39	1.6	60	2.7	125	5.0	369	82.5	599	6.2
Blue-collar	2,010	100	2,494	100	2,262	100	2,482	100	0	0	9,248	95.4
White-collar	0	0	0	0	0	0	0	0	447	100	447	4.6

Note. SPR = Supervisory Profile Record. Values are for all subjects rated by two raters.

therefore included in this validity generalization study (see discussion later). Thus, key development was conducted on over 10,000 supervisors who were then not included in the validity generalization (cross-validation) study. Table 1 shows the race, sex, and blue- versus white-collar composition of the developmental samples. Table 2 shows developmental sample values for age, education, years of company service, and supervisory experience.

Criteria

Two performance-rating criteria were (a) developed on the basis of a thorough job analysis conducted in all organizations of SPR I and (b) rechecked using the same job-analysis methods in all organizations in both developmental and validation consortia. Job analysis and instrument development were aimed at identifying and measuring the broad common core of supervisory tasks, and capacities to perform those tasks, shared across first-line supervisory jobs in all settings. The job-analysis results indicated that the basic duties and required capacities were very similar across organizations; intercorrelations of rated job requirements across organizations were in the .90s (Richardson, Bellows, Henry & Co., Inc., 1981). The resulting criterion instrument had two parts. Part I consisted of 28 statements about the individual's performance of specific job duties and an overall rating of performance across all duties. Ratings were made on 9-point scales, in terms of the "extent to which the individual meets job requirements." Anchors were *well below, somewhat below, meets, somewhat above,* and *well above* normal requirements. Part II contained 21 statements about the individual's specific supervisory abilities (e.g., ability to plan work for unit supervised) and a statement about overall ability to do the job. Ratings were made on a 9-point scale, similar to that for duty ratings, but in terms of "extent to which individual resembles other first-line supervisors." Anchors were identical to those for duty ratings, except that the mid-point of the scale was anchored by *average* rather than *meets.* Because supervisory jobs are similar but not identical in their duty and ability requirements, and some raters may not have had full opportunity to observe performance on some elements, raters were given the option on each statement to indicate that an element was not part of the job or that they could not evaluate the individual on that element.

Ratings were made by the immediate supervisor (Rater 1) of each individual in the sample and by an additional evaluator who knew the subject's performance well enough to rate it (Rater 2). Duty and ability composites consisted of the average rating across all rated elements. The duty and ability rating composites were then averaged separately across the two raters. The average of the two raters' average ratings, for duty and ability, respectively, served as the validation criteria. Reliability estimates for each of the criteria (the mean of two raters' ratings) were calculated by correlating the respective Rater 1 and Rater 2 averages and then adjusting this figure by using the Spearman–Brown formula for two raters. For the data used in our validity generalization study, this average reliability for the ability ratings was .69, and for the duty ratings, .64. The average interrater correlation between ability and duty ratings was .48 for one rater and .65 for averages based on two raters. These figures suggest a high degree of collinearity between the two ratings, and the two ratings correlated .98, on average, after correcting for unreliability in each rating. Data for both ratings, however, were analyzed separately, for three reasons. First, validities for the ability ratings averaged

Table 2
Developmental Sample Characteristics: Age, Education, Years of Service, and Supervisory Experience (Consortium Grouping A)

	SPR I			SPR II			SPR III			SPR IV[a]			SPR V		
Variable	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Age	1,979	39.62	10.08	2,490	42.91	10.24	2,258	42.52	9.97	2,476	42.99	10.42	446	35.42	9.95
Education[b]	1,998	3.28	1.77	2,468	3.23	1.75	2,242	3.14	1.93	2,469	3.52	1.81	444	3.53	1.66
Company service (years)	2,000	13.69	9.88	2,475	16.66	10.47	2,215	16.23	11.09	2,461	17.62	11.97	435	8.43	7.32
Supervisory experience (years)	NA			2,465	6.65	5.92	2,170	7.87	5.92	2,448	7.04	7.50	431	3.34	5.24

Note. Values are for all subjects rated by two raters. SPR = Supervisory Profile Record. NA = not available.
[a] Most of SPR IV is included in the developmental sample. However, a subgroup of SPR IV (N = 940) was not used in key development; these cases were included in the validation sample. [b] Education is coded (1) less than high school, (2) high school graduate, (3) high school graduate plus other formal non-college training, (4) less than 2 years of college, (5) 2 years of college, (6) 2 to 4 years of college, but did not graduate, and (7) college graduate.

Table 3
Validity Generalization Sample Characteristics: Race, Sex, and Collar (Consortium Grouping B)

	SPR IV[a]		SPR VI		SPR VII		SPR VIII		SPR IX		Grand total	
Sample	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Whites	904	96.3	4,081	94.5	1,600	86.4	2,313	93.8	1,600	91.0	10,498	92.7
Blacks	30	3.2	165	3.8	192	10.3	143	5.8	76	4.3	606	5.4
Hispanics	3	0.3	26	0.6	40	2.2	3	0.1	59	3.4	133	1.2
Other ethnic	2	0.2	31	0.7	19	1.0	8	0.3	24	1.4	84	0.6
Men	893	95.1	4,278	99.4	1,478	79.8	2,092	84.7	1,598	90.8	10,339	91.3
Women	46	4.9	25	0.6	373	20.2	377	15.3	161	9.2	989	8.7
Blue-collar	939	100	4,303	100	1,393	73.1	1,184	48.0	1,123	63.8	8,942	79.0
White-collar	0	0	0	0	458	24.8	1,285	52.0	636	36.2	2,379	21.0

Note. Values are for all subjects rated by two raters.
[a] Most of SPR IV is included in the developmental sample. However, a subgroup of SPR IV (N = 940) was not used in key development; these cases were included in the validation sample.

consistently higher than those for the duty ratings, even after correction for unreliability. Second, from a psychological point of view, the cognitive processes involved in assigning the two types of ratings appear to be different. Third, there is interest in the area of performance appraisal in the relative value and desirability of ratings based on tasks, duties, or behaviors versus ratings based on abilities. These questions are discussed in more detail in another article (Goff & Schmidt, 1988).

Scoring Key Development for the Biodata Subscale

Key development in SPR I. The prototype SPR tested on Consortium I contained 128 biodata items. Validation took place on the level of the individual item, that is, one or more alternative responses to each item had to be significantly related to the prediction of supervisory success to be keyed and retained. The entire sample of 2,010 supervisors in SPR I was split into two equal halves stratified to be equivalent in distribution of organizational membership and criterion ratings. This was done twice, once for duty ratings and once for ability ratings. Consider duty ratings first. *Successful* versus *unsuccessful* performance was evaluated in three nonindependent ways. In the first pass, *successful* was defined as having been rated at the criterion scale midpoint or above, whereas *unsuccessful* was defined as having been rated below the criterion scale midpoint. This procedure yielded a dichotomous criterion. The second variation defined *unsuccessful* as before, defined *successful* as the group at the top of the performance criterion equal in size to the group labeled *unsuccessful,* and assigned the remaining subjects to a middle group. This procedure yielded a three-value criterion scale. In the third and final performance grouping, the top and bottom portions of the middle group were split in such a way that the five resulting groups approximated a normal curve distribution of criterion performance. This procedure yielded a 5-point criterion scale. The use of three different variations of the duty-based criterion yielded three different (nonindependent) item analysis results for each biodata item for the duty ratings. The same procedure was followed for the ability ratings, creating a total of six sets of item analysis results. Items were retained and keyed (using unit weights) if (a) the item showed validity at the $p < .10$ level in four or more of the six item analyses and (b) the reason for the keyed relationship could be explained in logical, job-related terms (i.e., if there was a rationale or psychological explanation for the keyed relationship). Thus, although not as formal as some of the procedures suggested by Pace and Schoenfeldt (1977), this procedure ensured that the items passed a test of reasonableness as well as statistical significance. Items were also reviewed to see whether responses that were keyable in the total sample were similarly related to performance in the Black subsample. Responses not so related were re-evaluated, keeping in mind the Black sample's small size (N = 224). Finally, item weights for keyed items were summed for each individual to produce a total biodata score.

Key development in SPR II–V. The SPR biodata section was revised on the basis of results of the SPR I keying. Only keyed items and those promising enough to warrant further study were retained. This resulted in a revised version of the SPR, with 70 biodata items. The revised SPR was administered to the SPR II sample (N = 3,017; 10 organizations and 105 locations), and item analyses similar to those for SPR I were conducted, except that from this point on, only the five-group performance distribution was used as the criterion. These analyses were conducted on the SPR II sample and on the combined sample (SPR I plus SPR II), and the key was modified on the basis of the combined results (total N = 5,027 from 18 organizations). The revised key was tested on SPR III (N = 2,744; 8 organizations and 215 locations). This sample was used as a "tie-breaker," and moved the validation effort a step beyond the usual cross-validation model. With the addition of the SPR IV and V samples, the final total sample on which key development took place was increased to 10,697 supervisors from 39 organizations. The final key used in this study (the 1980 key) was based on evaluation of item analysis results for all five consortia comprising Consortium Grouping A; only items that were valid in all five consortia were retained in this key, which contains 41 items.

The Validity Generalization Research Sample

The independent validity generalization research sample consisted of Consortia VI, VII, VIII, and IX, plus a portion (N = 940; 38%) of Consortium IV. These groups combined are referred to as Consortium Grouping B; the total sample size is 11,332. Table 3 shows the race, sex, and white- versus blue-collar composition of Consortium Grouping B (i.e., the total validity generalization sample). Table 4 shows the validity generalization sample characteristics for age, education, years of service, and supervisory experience. Consortium Grouping B contains only purely cross-validational data; no changes whatsoever were made in the biodata key based on these data.

Only supervisors whose performance had been rated by two raters were included in the validity generalization sample; this step allowed determination of the reliability of ratings for each validity coefficient (see earlier Criteria section). Also, samples smaller than 10 were not included. Finally, individuals missing the relevant data code were omitted from that meta-analysis; for example, people whose coding on race (education) was missing were not included in the meta-analysis by race (educational level). Thus, the sample sizes in Tables 5 and 6 differ somewhat from those in Tables 3 and 4. For these same reasons, sample sizes varied slightly from one meta-analysis to another (see Tables 5 and 6).

Table 4
Validity Generalization Sample Characteristics: Age, Education, Years of Service, and Supervisory Experience (Consortium Grouping B)

Variable	SPR IV[a]			SPR VI			SPR VII			SPR VIII			SPR IX		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Age	936	43.01	10.24	4,285	46.71	9.24	1,850	42.06	10.54	2,463	43.87	10.01	1,750	45.88	9.07
Education[b]	934	3.68	1.84	4,285	3.03	1.49	1,844	3.50	1.85	2,464	3.95	1.94	1,657	2.74	1.29
Company service (years)	925	17.58	11.32	4,133	22.75	10.12	1,777	15.86	10.37	2,364	19.82	10.95	1,699	21.06	10.03
Supervisory experience (years)	916	7.29	6.54	4,073	7.59	6.34	1,732	6.48	6.50	2,309	7.20	6.51	1,686	8.82	6.92

Note. Values are for all subjects rated by two raters.
[a] Most of SPR IV is included in the developmental sample. However, a subgroup of SPR IV ($N = 940$) was not used in key development; these cases were included in the validation sample. [b] Education is coded (1) less than high school, (2) high school graduate, (3) high school graduate plus other formal non-college training, (4) less than 2 years of college, (5) 2 years of college, (6) 2 to 4 years of college, but did not graduate, and (7) college graduate.

The total sample size for the validity generalization analysis was 11,332 supervisors whose performance had been rated by two raters. All validity data used in our study are concurrent in nature. Both white-collar and blue-collar supervisors were included, and the organizations represented came from a variety of different industries (utilities, petroleum, automotive, communications, steel, chemicals, banking, food processing, and several others).

Potential Moderators

Although the focal analysis in this research is on the organization as a moderator of biodata validities, the literature contains hypotheses about other potential moderators, as noted earlier. Thayer (1977), for example, has suggested that age, sex, race/ethnicity, and prior experience could moderate biodata validities; Reilly and Chao (1982) have concluded that different keys may be necessary for men and women; and both Sparks (1983) and Owens (1976) have suggested the necessity of removing the effects of age and experience from biodata validities. To examine whether the validity of the single SPR key generalizes across these variables, separate meta-analyses were conducted by age, sex, race, education, tenure, and supervisory experience.

For race, within each consortium separate validities were computed for Whites, Blacks and Hispanics; these validities were then entered into the meta-analysis for race. The same procedure was followed for sex and for the white- versus blue-collar variable. Education was coded into seven levels (see Footnote c to Table 4). Data were pooled and a separate validity was computed for each of these seven educational levels. Company service, supervisory experience, and age were all coded in years; a validity was computed separately for each year-code. One validity was computed for each age, each number of years of service, and for each number of years of supervisory experience. No validity coefficient was computed if there were less than 10 people in a cell.

Meta-Analysis Method Used

In the data set used in this study, complete artifact information was available for each correlation. That is, for each observed validity coefficient, there was an associated criterion reliability and index of range restriction. (Predictor reliability was also known, but because the predictor was always the same scale, the reliability was constant across studies; hence, there was no need to correct for the effects of differences between studies in predictor reliability; see Schmidt, Hunter, Pearlman, & Hirsh, 1985, Q&A 31). Therefore, it was possible to (a) correct each observed validity individually for criterion unreliability and range restriction, and (b) perform the meta-analysis on the corrected correlations. The final correction in this form of meta-analysis is the correc-

Table 5
Combined Validity Generalization Results for Criteria of Ability to Perform the Job

Variable	*N*	No. of *rs*	Mean observed	Observed SD_{r_c}	Predicted SD_{r_c}	Percentage variance accounted for	$\hat{\rho}$	$SD_{\hat{\rho}}$	90% CV
Organization	11,288	79	.30	.126	.095	58	.36	.082	.26
Race (by consortium)	11,229	13	.29	.037	.036	96	.34	.007	.33
Sex (by consortium)	11,321	10	.28	.046	.032	48	.32	.033	.28
Education	11,184	7	.28	.039	.028	49	.34	.028	.30
Collar (by consortium)	11,321	8	.28	.030	.028	92	.33	.010	.32
Company service (years)	10,885	45	.28	.075	.070	87	.33	.027	.30
Supervisory experience (years)	10,220	22	.28	.040	.050	156	.33	.000	.33
Age	11,332	43	.27	.054	.069	165	.32	.000	.32
Means			.28	.056	.043	77[a]	.33	.023	.30

Note. CV = credibility values.
[a] An unbiased estimate of mean percentage of variance accounted for across meta-analyses, calculated by taking the reciprocal of the average of reciprocals of individual predicted to observed variance ratios. See text for details.

Table 6
Combined Validity Generalization Results for Criteria of Performance on Job Duties

Variable	N	No. of rs	Mean observed	Observed SD_r	Predicted SD_r	Percentage variance accounted for	$\hat{\rho}$	$SD_{\hat{\rho}}$	90% CV
Organization	11,288	79	.27	.144	.100	48	.34	.104	.20
Race (by consortium)	11,229	13	.27	.043	.038	80	.32	.019	.30
Sex (by consortium)	11,321	10	.26	.043	.033	59	.31	.027	.29
Education	11,184	7	.26	.034	.028	68	.32	.019	.29
Collar (by consortium)	11,321	8	.26	.030	.030	99	.31	.003	.31
Company service (years)	10,885	45	.26	.074	.073	96	.31	.014	.30
Supervisory experience (years)	10,220	22	.26	.052	.066	161	.31	.000	.31
Age	11,332	43	.25	.056	.069	161	.30	.000	.30
Means			.26	.060	.049	82[a]	.32	.023	.29

Note. CV = credibility values.
[a] An unbiased estimate of mean percentage of variance accounted for across meta-analyses, calculated by taking the reciprocal of the average of reciprocals of individual predicted to observed variance ratios. See text for details.

tion for the sampling error variance of the *corrected* correlations. However, note that the earlier correction of each observed validity for criterion unreliability and range restriction corrects for validity differences between studies due to variations in these two artifacts. This method of meta-analysis is more exact than the more commonly used method that uses distributions of artifacts, the values of which are not associated with specific observed coefficients. Further details are given in Hunter, Schmidt, and Jackson (1982, chapter 3) and in Hunter and Schmidt (in press, chapter 4).

Although individual correction and subsequent analysis of corrected correlations is the preferred method of meta-analysis, it has not yet been applied in a published study because the necessary information has generally not been available. Two unpublished dissertations based on military data (Pearlman, 1982; Stern, 1987) used this method. Brown (1981) did correct each coefficient individually and computed the mean of the corrected correlations; however, his variance corrections were made by using artifact distribution-based meta-analysis (although these distributions were derived from the data set analyzed). At that time there were no published descriptions of methods of meta-analysis for individually corrected correlations. Thus, the present meta-analysis is methodologically unique: (a) It comes from a major civilian job family in a large number of organizations, (b) it contains complete artifact information on every validity coefficient, (c) it corrects each coefficient individually, and (d) the full meta-analysis is performed on the corrected correlations.

Criterion reliabilities were computed as explained earlier. The corrections for criterion unreliability were made before the corrections for range restriction because the reliabilities were computed directly on the groups being studied (Hunter et al., 1982); that is, the criterion reliability estimates were for the restricted group. The applicant standard deviation was 4.72, a value that is almost the same as the average incumbent (restricted) value (mean restricted SD = 4.68). The applicant standard deviation was based on 17,962 recent candidates for promotion to first-line supervisor in a number of organizations using the SPR as part of the selection process. Candidates are typically nominated by their supervisors and must successfully meet several other prescreening requirements before being allowed to take the SPR. Thus, candidates are a somewhat homogeneous group, but these processes accurately reflect how applicant pools are created in organizations. The incumbents in the validity studies, on the other hand, were often older, less educated, and entered their supervisory positions at a time when selection standards were lower. Thus, they are typically about as variable on the bio-data scale as current applicants. Because of this, the correction for range restriction in these analyses had little or no effect on the resulting mean validity estimates.

In each meta-analysis, the mean sample size weighted validity was computed, and the variance across the correlations was calculated weighting each validity by its sample size. The amount of variance predicted on the basis of sampling error, for frequency weighted corrected correlations, was computed on the basis of the following formula (Hunter et al., 1982, p. 71):

$$\sigma_{e_c}^2 = \frac{1}{N}\sum N_i \alpha_i^2 \frac{(1-\bar{r}^2)^2}{N_i - 1}$$

where

$\sigma_{e_c}^2$ = the sampling error variance of the distribution of corrected correlations,
N = total sample size,
N_i = sample size for each group,
$\bar{r}$ = the mean *uncorrected* correlation,

and

$$\alpha = \left(\frac{r_c}{r}\right)\left(\frac{1}{(U^2-1)r^2+1}\right),$$

where r_c = the corrected correlation, r = the uncorrected correlation, and $U = S/s$, the ratio of the unrestricted to the restricted predictor standard deviation. Further explanation is given in the Appendix.

Results

Ability Ratings

Results using the average ability composite rating as the criterion are presented in Table 5. The first four columns of numbers in Table 5 contain the total sample size, the number of validity coefficients on which each distribution was based, the uncorrected (i.e., observed) mean validity, and the standard deviation (SD) of the *corrected* validities, respectively. The predicted SD is the standard deviation of the corrected validities that would be predicted on the basis of the sampling error in the (individually) corrected validities. The next column reports

the percentage of observed variance that is accounted for by sampling error (i.e. the ratio of sampling error variance to observed variance).[1] Callender and Osburn (1988) showed that the arithmetic average of the percentage of variance accounted for across meta-analyses produces an overestimate. This results from the fact that the *average* ratio of expected (from artifacts) variances to observed variance has an upward bias (stemming from the fact that observed variance is sometimes, by chance, extremely small). However, the reciprocal average ratio, the average of the ratio of observed to expected variance, is not biased. To eliminate this bias, the reciprocal of each percentage of variance accounted for (i.e., the observed to error variance ratios) was obtained, the reciprocals were averaged, and the reciprocal of the average was obtained. The latter value is the unbiased estimate of the average percentage of variance accounted for across all meta-analyses presented in the table (Hunter & Schmidt, in press). The last three columns present the mean, standard deviation, and 90% credibility values, respectively, for the estimated true validity distributions.

The magnitude of the average observed mean validities is similar to, but perhaps slightly smaller than, those reported in previous reviews of biodata validity. The observed standard deviations are all relatively small, as would be expected on the basis of the large average sample sizes and the use of the same predictor and criterion measure in all the studies. In five of the eight meta-analyses, over 75% of the observed variance of the corrected correlations is accounted for by sampling error. The unbiased estimate of average percentage of variance accounted for across all the meta-analyses is 77%. The average observed standard deviation of the corrected validities is .056, and the average predicted from artifacts is .043, for a mean difference of only .013. These findings suggest there is little true variation in the validities in these data.

A key outcome in any validity generalization analysis is SD_ρ, the estimated standard deviation of true validities. In Table 5, two of the eight values for SD_ρ are zero and six are greater than zero. To focus ex post facto on only the SD_ρ values greater than zero is to capitalize on chance (i.e., on second-order sampling error). However, if we nevertheless examine those specific SD_ρs that have numerical values greater than zero, we find that they are relatively small. Table 5 shows that the estimated true standard deviations are greater than zero for the following meta-analyses: organization (.082), race (.007), sex (.033), collar (.010), years of company service (.027), and education (.028). Thus, direct examination of the non-zero estimated true *SD*s (SD_ρs), as well as the more general analysis previously discussed, leads to the conclusion that the amount of variation remaining after the validities are corrected for artifacts is relatively small, and provides little support for moderator hypotheses.

Additional documentation of the generalizability of biodata validities across the potential moderators examined here is given by the similarity between the estimated mean true validities and the 90% credibility values. For all potential moderators except organizations, the 90% credibility values are quite similar to the mean estimated true validities, with mean estimated true validities ranging from .32 to .34, and 90% credibility values between .28 and .33. Though still relatively small, the SD_ρ value for organizations is larger than for the other potential moderators. However, this SD_ρ value of .082 is considerably smaller than the corresponding value obtained by Schmidt et al. (1979, Table 6) for general mental ability for the same job (first-line supervisors). In our study, the estimated mean true validity for organizations is .36, and the 90% credibility value is .26. Thus, across organizations, 90% of all biodata validities based on the SPR are expected to be at least .26. Given the larger SD_ρ value, the probability that organizations exert some moderating effect on biodata validities is perhaps greater than in the case of the other potential moderators examined. However, this probability must still be viewed as very small. SD_ρ values larger than .082 have been found in the cognitive domain to be fully consistent with the conclusion that there is no real variance in true validities (Schmidt, Hunter, Pearlman, & Hirsh, 1985). As is virtually always the case, our meta-analysis was not able to correct for all sources of artifactual variance. For this reason, the remaining variance in a meta-analysis, especially if small (as in the case here) must always be interpreted cautiously (Hunter & Schmidt, in press; Schmidt, Hunter, Pearlman, & Hirsh, 1985).

Duty Ratings

Table 6 displays the results of meta-analyses based on duty ratings as the criterion. They are similar to those obtained for ability ratings, with only a few differences. The observed validities are all slightly smaller than those for ability (by approximately .02), and the observed standard deviations are slightly (.01) higher. Predicted *SD*s are also slightly higher, reflecting the effect of the slightly lower average reliability of the duty ratings. Percentage of variance accounted for does not vary uniformly in one direction or another. The average percentage of variance accounted for across all 8 meta-analyses is 82%. This corresponds to a difference of only .014 between the mean predicted and observed *SD*s of the validities (.049 predicted vs. .063 observed). The estimated mean true validities based on duty rat-

[1] A reviewer was disturbed by the fact that some of these percentages were considerably greater than 100%. The largest values in Table 5 are 156% and 165%. If the true value is 100%, values greater than 100% are expected 50% of the time due to second order sampling error (Callender & Osburn, 1988; Hunter & Schmidt, in press, chapter 9; Schmidt, Hunter, Pearlman, & Hirsh, 1985, Q&A 25). The reviewer acknowledged this but felt the figures should not go *that much* over 100%. But this is a common and expected result. Since the observed variance is a very tiny number, very small increments of predicted (from artifacts) variance over observed variance lead to percentages considerably over 100%. For example, in the case in Table 5 in which the figure is 165%, the predicted variance is only .0009 greater than the observed variance. Also, it is important to note that second order sampling error is caused by instability in the *observed* variance of correlations, not in the predicted variance. Observed variance can vary widely across identical data conditions simply because of chance (i.e., second order sampling error). In situations in which all variance is in fact due to artifacts, whenever observed variance happens to be smaller than its average (or expected) value (i.e., about 50% of the time), then the variance accounted for figure will go above 100%. This figure will go below 100% equally often. For further discussion of this point, see Schmidt, Hunter, Pearlman, & Hirsh (1985, Q&A 25), Callender and Osburn (1988), and Hunter and Schmidt (in press, chapter 9).

ings are an average of .01 lower than those for ability ratings, as is the average 90% credibility value. It is apparent that the conclusions about the generalizability of SPR validities across organizations, sex, race, age, education, consortium, collar, tenure, and experience are essentially the same, whether based on duty ratings or ability ratings as the criterion.

Discussion

The results presented here contraindicate the prevailing belief that biodata validities are intrinsically specific to particular organizations. They also present strong counterevidence to the hypothesis that biodata validities are necessarily moderated by age, sex, race, education, tenure, or previous experience. The 90% credibility values are on average only .03 lower than the mean true validity, providing strong support for validity generalization; the magnitudes of these values (approximately .33) indicate that the generalized validity is substantial. The current results do not, of course, indicate that the level of generalizability demonstrated here can always be expected from biodata; given conventional methods of biodata instrument construction and validation, these results may represent the exception rather than the rule. The point is that biodata instruments can be constructed and validated in a way that will lead to validity generalizability. The findings in this study show that large sample sizes, multiple organizations, and cross-organizational keying of the biodata scale can yield generalizable validities. Thus, the current findings also point up the advantages of consortium-based, multiple-organization biodata research.

The advantages of generalizable biodata validities are obvious. A relevant question is whether organizationally specific keys, although having the disadvantage of nongeneralizable validities, have higher validities for the specific organization for which they are developed. This question should be addressed in future research. Future researchers might also examine the *degree* of variability of validity across organizations of biodata scales developed in a single organization. However, most organizations do not have the necessary large samples to develop their own biodata scale, and thus generalizably valid biodata scales are the only ones they can use.

This research has not directly examined the question of temporal specificity or stability. Although there is evidence that carefully constructed biodata questionnaires retain their factor structure over time (Lautenschlager & Shaffer, 1987), it is a common finding that the validity of a specific biodata key decays over time, resulting in a requirement for rekeying (Hunter & Hunter, 1984; Thayer, 1977). Our study can, however, offer some indirect evidence bearing on this issue. Because the data from the first consortium were gathered in 1974 and the data in SPR IX were gathered in 1985, a time span of over 10 years was covered. This was a time span in which many social changes were taking place. Nevertheless, a key composed of items keyed in the developmental samples yielded substantial validities in the cross-validation sample, up to 11 years later. It may be that methods of biodata scale construction and validation based on large samples and successive replications produce both validities that generalize across organizations, and across other potential moderator variables, *and* validities that tend to be stable over fairly long periods of time. That is, generalizability and temporal stability of biodata validities may both depend on the same processes of scale construction. Future research is needed on this question.

The findings of this study bear on another hypothesis important to the use of biodata in selection: the hypothesis that biodata validity in concurrent studies stems from the measurement of knowledge acquired from job experience. If this hypothesis were correct, then concurrent validities would typically be much larger than the corresponding predictive validities. This would be a serious problem, because predictive validities are the basis for selection utility. If this hypothesis were correct, then concurrent validities should be much smaller when experience on the job is held constant. That is, this hypothesis holds that it is differences between individuals in job experience that create individual differences in biodata scores and therefore cause the validity of biodata scores. In the meta-analyses of years of supervisory experience, each validity coefficient was computed on individuals with the same number of years of experience on the job; that is, job experience was held nearly constant. Yet, mean validity did not decline for either ability (Table 5) or duty (Table 6) ratings. Thus, biodata validity does not appear to be an artifact of individual differences in job experience.

One reviewer maintained that because we did not include a group with zero experience, we could not conclude that knowledge acquired on the job does not influence biodata validity. It is, of course, logically impossible to include a group with zero job experience, but the reviewer's hypothesis was that minimal levels of job experience may be sufficient to "create" validity. To address this hypothesis, we examined validity separately for those supervisors with 3 years or less job experience (N = 3,611). In this group, a validity was computed for each 3-month interval of job experience. A meta-analysis of these validities showed that among these relatively less experienced supervisors, mean true validity was .31 for the ability ratings and .30 for the duty ratings. The corresponding values for all supervisors are .33 for ability ratings (Table 5) and .31 for duty ratings (Table 6), very similar values. In addition, the true validities for supervisors with only 1 month's job experience (N = 51) were .30 for ability ratings and .26 for duty ratings. Thus, validities do not appear to be affected by job experience. If there is an effect, it would appear to occur within the first month on the job, and early in that period at that. We judge this to be unlikely.

The question of whether biodata validities are created by individual differences in job experience is not the same as the question of whether concurrent validities are the same as predictive validities. All validities in our study were concurrent. The job experience hypothesis, which these data contraindicate, is only one possible hypothesis that might predict differences between concurrent and predictive biodata validities. For example, greater response distortion on the part of job applicants is another such hypothesis. There is a substantial amount of empirical evidence indicating that for cognitive tests, concurrent and predictive validities are very similar. However, we cannot assume that these findings for measures of maximum performance will also hold true for measures of typical performance, such as biodata scales. We believe, therefore, that research comparing predictive and concurrent validities for the same biodata scale would be useful.

Finally, the findings of this study bear on an hypothesis that is important for personnel selection in general: the hypothesis of situational specificity of validities (Schmidt, Hunter, Pearlman, & Hirsh, 1985). In many individual validity generalization studies, the findings indicate that the presence of validity can be generalized, but statistical artifacts do not appear to explain all the variability in validities. In the case of cognitive ability tests, substantial evidence has been presented that the remaining variability in validities is also due to artifacts, artifacts that cannot be corrected for (Schmidt & Hunter, 1984; Schmidt, Hunter, Pearlman, & Hirsh, 1985; Schmidt, Ocasio, Hillery, & Hunter, 1985). Some researchers have maintained that in the case of noncognitive predictors, such as biodata scales, this remaining amount of validity variability can be expected to be considerably larger than in the case of cognitive ability tests (Sackett, Schmitt, Tenopyr, Kehoe, & Zedeck, 1985; Schmitt & Schneider, 1983). In the case of noncognitive predictors, it is expected that such factors as organizational value systems, management philosophies, leadership styles, and organizational cultures will be major determinants of what kinds of people are successful in the organization (Schmitt & Schneider, 1983), and thus will be major moderators of the validities of noncognitive selection procedures. This moderating effect is expected to be much larger than in the case of cognitive abilities; thus, noncognitive predictors become a critical test for the situational specificity hypothesis.

The findings of this study contraindicate this hypothesis for one type of noncognitive predictor: biodata scales constructed using the methods described in this study. In this study, the standard deviation of biodata validities across organizations was quite small (.082 for ability ratings and .104 for duty ratings), limiting the room in which situation moderators could operate. This finding suggests that high levels of situational specificity are not an inherent property of noncognitive measures.

From a psychological point of view, our findings indicate that biodata questionnaires are capable of capturing general characteristics of people that conduce to success or failure on the job in a wide variety of settings, organizational climates, technologies, and so on. This has been the premise of Owens's research, particularly his assessment–classification model (Owens, 1968, 1976; Owens & Schoenfeldt, 1979; Schoenfeldt, 1974). This study has focused on the major job family of first-line supervisors; similar studies are now under way for managers and clerical personnel. At some point, research should be planned that explicates the psychological meaning of biodata items predictive of success at work. Although difficult to conduct because of the test security requirements for scoring keys, such research could be of considerable value to those interested in understanding the relationship between life history and work performance.

In summary, this research has shown that the validity of a well-developed autobiographical questionnaire instrument generalizes across a major job family: first-line supervision. All biodata items were based on a review of information about the job. Each item was based on a rationale or hypothesis as to its applicability to the candidate population; no item was keyed unless the relationship could be explained in psychological terms and the item showed validity across different organizations. All developmental samples were large, and the stability of all relationships was determined through later replications in multiple organizations. The findings of this study indicate that the validity of this instrument is temporally stable as well as generalizable. The findings also provide evidence against the hypothesis that the validity of biodata stems from measurement of knowledge and skills acquired on the job. Finally, the results of this study constitute additional evidence against the general hypothesis of situational specificity of validities; the findings disconfirm that hypothesis in an important noncognitive domain. This is significant because it has been hypothesized that situational specificity can be expected to be greater in noncognitive than in cognitive domains.

References

Asher, J. J. (1972). The biographical item: Can it be improved? *Personnel Psychology, 25,* 251–269.

Brown, S. H. (1981). Validity generalization in the life insurance industry. *Journal of Applied Psychology, 66,* 664–670.

Callender, J. C., & Osburn, H. G. (1988). Unbiased estimation of sampling variance of correlations. *Journal of Applied Psychology, 73,* 312–315.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance and effectiveness.* New York: McGraw-Hill.

Dreher, G. F., & Sackett, P. R. (1983). *Perspectives on staffing and selection.* Homewood, IL: Irwin.

Dunnette, M. D., Rosse, R., Houston, J. S., Hough, L. M., Toquam, J., Lammlein, S., Bosshardt, M. J., & Keyes, M. (1982). *Development and validation of an industry-wide electric power plant operator selection system.* Minneapolis, MN: Personnel Decisions Research Institute.

Goff, S., & Schmidt, F. L. (1988). *Task-based vs. ability-based ratings of job performance: Is the conventional wisdom wrong again?* Unpublished paper, Department of Industrial Relations and Human Resources, University of Iowa.

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

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U.S. Employment Service, Department of Labor.

Hunter, J. E., & Hunter, R. F. (1984). The validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–99.

Hunter, J. E., & Schmidt, F. L. (in press). *Modern meta-analysis.* Beverly Hills, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Laurent, H. (1970). Cross-cultural validation of empirically validated tests. *Journal of Applied Psychology, 54,* 417–423.

Lautenschlager, G. J., & Shaffer, G. S. (1987). Reexamining the component stability of Owens's biographical questionnaire. *Journal of Applied Psychology, 72,* 149–152.

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

Owens, W. A. (1968). Toward one discipline of scientific psychology. *American Psychologist, 23,* 782–785.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609–644). Chicago: Rand-McNally.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology, 63,* 569–607.

Pace, L. A., & Schoenfeldt, L. F. (1977). Legal concerns in the use of weighted applications. *Personnel Psychology, 30,* 159–166.

Pearlman, K. (1982). *The Bayesian approach to validity generalization: An examination of the robustness of procedures and conclusions.* Unpublished doctoral dissertation, George Washington University.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–62.

Richardson, Bellows, Henry & Co., Inc. (1981). Supervisory Profile Record (Technical Reps., Vols. 1, 2, & 3). Washington, DC: Author.

Sackett, P. R., Schmitt, N., Tenopyr, M. L., Kehoe, J., & Zedeck, S. (1985). Commentary on forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697–798.

Schmidt, F. L., & Hunter, J. E. (1984). A within-setting testing of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 37,* 317–326.

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

Schmitt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32,* 257–281.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38,* 509–524.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsh, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Schmitt, N., & Schneider, B. (1983). Current issues in personnel selection. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 1). Greenwich, CT: JAI Press.

Schoenfeldt, L. F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology, 59,* 583–594.

Sparks, C. P. (1983). Paper and pencil measure of potential. In G. F. Dreher & P. R. Sackett (Eds.), *Perspectives on staffing and selection.* Homewood, IL: Irwin.

Stern, B. (1987). *Job complexity as a moderator of the validity of the Armed Services Vocational Aptitude Battery.* Unpublished doctoral dissertation, George Washington University.

Thayer, P. W. (1977). Somethings old, somethings new. *Personnel Psychology, 30,* 513–524.

Appendix

Improved Method for Estimation of Sampling Error of Individually Corrected Correlation Coefficients

As noted by Hunter et al. (1982), the sampling error of correlation coefficients corrected for measurement error or range restriction, or both, is larger than that of the uncorrected coefficients. These authors presented an estimate of the sampling error of a corrected coefficient, that is, a good approximation when the observed coefficient is small, when the U value does not differ too severely from 1.0. U is the ratio of the standard deviation in the reference population to the study standard deviation. When U deviates considerably from 1.00, the approximation presented by Hunter et al. becomes less accurate because the estimation formula assumes linear transformation of the observed correlation coefficient, whereas the range restriction correction is nonlinear.

In our study, coefficients were first corrected for measurement error and then for range restriction or enhancement. As in Hunter et al. (1982), the estimated sampling error for an individually corrected coefficient is

$$\sigma^2_{e_c} = \alpha^2 \sigma^2_e,$$

where

$\sigma^2_{e_c}$ = the sampling error of the corrected coefficient,
α^2 = the squared multiplier (defined later), and
σ^2_e = the sampling error of the uncorrected coefficient.

The new estimation formula entails a more accurate calculation of α, as follows:

$$\alpha = \left(\frac{1}{((U^2 - 1)r^2) + 1}\right)\left(\frac{r_c}{r}\right),$$

where

α = the multiplier,
r = the observed coefficient,
r_c = the coefficient corrected for range restriction or enhancement and for measurement error and range restriction or enhancement, and
U = the ratio of the population standard deviation to the sample standard deviation.

The second term in this equation is α, as given in Hunter et al. (1982; chapter 3). The first term reflects the nonlinearity in the range correction; this term is less than one, and its omission causes a slight overestimation of the amount of variance in corrected coefficients that is due to sampling error (Hunter & Schmidt, in press, chapter 3).

Received March 17, 1988
Revision received October 23, 1989
Accepted October 24, 1989 ■