Journal of Applied Psychology
2003, Vol. 88, No. 4, 635–646

0021-9010/03/$12.00 DOI: 10.1037/0021-9010.88.4.635

A Meta-Analysis of Job Analysis Reliability

Erich C. Dierdorff and Mark A. Wilson
North Carolina State University

Average levels of interrater and intrarater reliability for job analysis data were investigated using meta-analysis. Forty-six studies and 299 estimates of reliability were cumulated. Data were categorized by specificity (generalized work activity or task data), source (incumbents, analysts, or technical experts), and descriptive scale (frequency, importance, difficulty, time-spent, and the Position Analysis Questionnaire). Task data initially produced higher estimates of interrater reliability than generalized work activity data and lower estimates of intrarater reliability. When estimates were corrected for scale length and number of raters by using the Spearman-Brown formula, task data had higher interrater and intrarater reliabilities. Incumbents displayed the lowest reliabilities. Scales of frequency and importance were the most reliable. Implications of these reliability levels for job analysis practice are discussed.

Since mandating the legal requirements for the use of job analyses (Uniform Guidelines for Employee Selection Procedures, 1978), the importance of obtaining job analysis data and assessing the reliability of such data has become a salient issue to both practitioners and researchers. It has been estimated that large organizations spend between $150,000 and $4,000,000 annually on job analyses (Levine, Sistrunk, McNutt, & Gael, 1988). Furthermore, it appears probable that job analysis data will continue to undergo increasing legal scrutiny regarding issues of quality, similar to the job-relatedness of performance appraisal data, which have already seen a barrage of court decisions during the past decade (Gutman, 1993; Werner & Bolino, 1997). Considering the widespread utility implications, legal issues, and organizational costs associated with conducting a job analysis, it would seem safe to assume that the determination of the general expected level of job analysis data reliability should be of primary importance to any user of this type of work information.

Prior literature has lamented the paucity of systematic research investigating reliability issues in job analysis (Harvey, 1991; Harvey & Wilson, 2000; Morgeson & Campion, 1997; Ployhart, Schmitt, & Rogg, 2000). Most research delving into the reliability and validity of job analysis has been in a search for moderating variables of individual characteristics, such as demographic variables like sex, race, or tenure (e.g., Borman, Dorsey, & Ackerman, 1992; Landy & Vasey, 1991; Richmann & Quinones, 1996) or other variables like performance and cognitive ability (e.g., Aamodt, Kimbrough, Keller, & Crawford, 1982; Harvey, Friedman, Hakel, & Cornelius, 1988; Henry & Morris, 2000). The overall conclusions of these research veins have been mixed, with some showing significant evidence of moderation, and others displaying none. It is interesting to note that only recently have the definitions of reliable and valid job information received directed attention and discourse (Harvey & Wilson, 2000; Morgeson & Campion, 2000; Sanchez & Levine, 2000). More recent research has tended to frame the quality of job analysis data through views ranging from various validity issues (Pine, 1995; Sanchez & Levine, 1994), to potential social and cognitive sources of inaccuracy (Henry & Morris, 2000; Morgeson & Campion, 1997), to the merits of job analysis and consequential validity (Sanchez & Levine, 2000), and to an integrative approach emphasizing both reliability and validity examinations (Harvey & Wilson, 2000; Wilson, 1997). As an important component of data quality, we sought to specifically examine the role of reliability in relation to job analysis data quality.

Purpose

The principal purpose of this study was to provide insight into the average levels of reliability that one could expect of job analysis data. Coinciding with this purpose were more specific examinations of the reliability expectations given different data specificity, various sources of data, variety of descriptive scales, and techniques of reliability estimation. The hope embedded in estimating average levels of reliability was that these data may in turn inspire greater attention to the reliability of job analysis data, as well as be used as reference points when examining the reliability of such data. We feel that not enough empirical attention has been paid to this issue, and that the availability of such reliability reference points could be of particular importance to practitioners conducting job analyses. To date, no such estimates have been available, and practitioners have had no means of comparison with which to associate the reliability levels they may have obtained. Moreover, elucidation of the levels of reliability across varying data specificity, data sources, and descriptive scales would provide useful information regarding decisions surrounding the method, sample, format, and overall design of a job analysis project.

Erich C. Dierdorff and Mark A. Wilson, Department of Psychology, North Carolina State University.

This research was conducted in partial fulfillment of the requirements for the master's degree at the North Carolina State University by Erich C. Dierdorff.

Correspondence concerning this article should be addressed to Mark A. Wilson, Department of Psychology, North Carolina State University, P.O. Box 7801, Raleigh, North Carolina 27695-7801. E-mail: mark_wilson@ncsu.edu

Scope and Classifications

Work information may range from attributes of the work performed to the required attributes of the workers themselves. Unfortunately, this common collective conception of work information (job-oriented vs. worker-oriented) can confound two distinctive realms of data. Historically, Dunnette (1976) described these realms as "two worlds of human behavioral taxonomies" (p. 477). Dunnette's two worlds referred to the *activities required* by the job and the *characteristics of the worker* deemed necessary for successful performance of the job. More recently, Harvey and Wilson (2000) contrasted "job analysis" versus "job specification," with the former collecting data about work activities and the latter collecting data describing worker attributes presumably required for job performance. The present study focused only on reliability evidence obtained through data that described the activities performed within a given work role (i.e., job analysis). This parameter allowed the study's investigations to examine the reliability of data that carry the feasibility of verification through observation, as opposed to latent worker attributes typically described by job specification data.

The primary classification employed by the present study delineated job analysis data by two categories of specificity: task and general work activity (GWA). These classifications were not meant to be all-inclusive but rather were meant to capture the majority of job analysis data. Task-level data were defined as information that targets the more microdata specificity (e.g., "cleans teeth using a water-pick" or "recommends medication treatment schedule to patient"). In contrast, GWA-level data were defined similarly to the description offered by Cunningham, Drewes, and Powell (1995), portraying GWAs as "generic descriptors," including general activity statements applicable across a range of jobs and occupations (e.g., "estimating quantity" or "supervising the work of others"). An important caveat to data inclusion was that only GWAs relating to the work performed within a job were used, thus excluding what have been referred to as "generalized worker requirements" such as knowledge, skills, and abilities (KSAs; McCormick, Jeanneret, & Mecham, 1972). By separately coding tasks and GWAs, the present study allowed an investigation of job-analysis reliability relative to the specificity domain of the collected data. Prior literature has suggested that job-analysis data specificity may affect the reliability of such data (Harvey & Wilson, 2000; K. F. Murphy & Wilson, 1997), with more specific data showing higher reliability levels. Moreover, with the increasing prevalence of "competency" modeling, which in part incorporates more general levels of behavioral information (Schippmann, 1999), as well as the recent push to more generic activities for purposes of job and occupation analysis (Cunningham, 1996), the separate examination of GWAs allowed for interpretative comparisons with increasingly prevalent and contemporary job and occupation analysis approaches.

In addition to data specificity, the present study incorporated a classification for the source from which the data were generated. Sources of job-analysis information were classified into three groupings: (1) incumbents, (2) analysts, and (3) technical experts. Incumbent sources referred to job information derived from jobholders. These data were usually collected through self-report surveys and inventories. Analyst derived job information was from nonjobholder professional job analysts. These data were generally gathered through methods such as observation and interviewing and were then used to complete a formal job-analysis instrument (i.e., Position Analysis Questionnaire [PAQ]). The third source group, technical experts, captured data obtained through individuals defined specifically as training specialists, supervisors, or higher level managers (Landy & Vasey, 1991). Because many technical experts can also be considered job incumbents, this designation was reserved only for data that were explicitly described as being collected from technical experts, supervisors, or some other "senior level" source. By source-coding reliability evidence, analyses could reveal any changes in the magnitude of reliability estimates in relation to these common sources of job-analysis data. Prior empirical investigation has suggested the possibility of differential levels of reliability across various classifications of respondents (Green & Stutzman, 1986; Henry & Morris, 2000; Landy & Vasey, 1991), such as performance level of the incumbent and various demographic characteristics of subject matter experts. The present research sought to compare the reliability levels across sources rather than only within a given source as in previous research.

A third classification was used to categorize the type of descriptive scale upon which a job was analyzed. Some common examples of descriptive scales are time spent on task, task importance, and task difficulty (Gael, 1983). Past research has suggested that the variety of scales used in job analysis yield different average reliability coefficients (Birt, 1968). For instance, scales of frequency of task performance and task duration have displayed reliabilities ranging from the .50s to the .70s (McCormick & Ammerman, 1960; Morsh, 1964). Difficulty scales have generally been found to have lower reliabilities than other descriptive scales, with estimates ranging from the .30s to the .50s (McCormick & Ammerman, 1960; McCormick & Tombrink, 1960; Wilson, Harvey, & Macy, 1990). Thus, data were coded for the commonly used descriptive scales of frequency, importance, difficulty, and time spent. GWA data derived from the PAQ (McCormick, Jeanneret, & Mecham, 1972), which is arguably the most widely used and researched generic job analysis instrument, were additionally coded.

To allow for a comparative analysis of reliability across the three aforementioned classifications, it was necessary to group coefficients into appropriate estimation categories. Therefore, reliability estimates were delineated by their computational approach. Two approaches commonly used in job analyses to estimate reliability were chosen as the categories employed by this study. Both types of reliability estimation are discussed in the ensuing section.

Types of Reliability Estimates Used in Job Analysis

The two most commonly used forms of reliability estimation are interrater and intrarater reliability (Viswesvaran, Ones, & Schmidt, 1996). In the context of job analysis practice, interrater reliability seems to be the more prevalent of the two techniques. Interrater reliability identifies the degree to which different raters (i.e., incumbents) agree on the components of a target work role or job. Interrater reliability estimations are essentially indices of rater covariation. This type of estimate can portray the overall level of consistency among the sample raters involved in the job analysis effort. Typically, interrater reliability is assessed using either Pear-

son correlations or intraclass correlations (ICC; see Shrout & Fleiss, 1979, for a detailed discussion). Most previous empirical literature has focused on the intrarater reliability of job analysis data. Two forms of intrarater reliability commonly employed within job analysis are repeated item and rate–rerate of the same job at different points in time. Both of these estimates may be viewed as coefficients of stability (Viswesvaran et al., 1996). The repeated items approach can display the consistency of a rater across a particular job analysis instrument (i.e., task inventory), whereas the rate–rerate technique assesses the extent to which there is consistency across two administrations. Intrarater reliability is typically assessed using Pearson correlations.

Research Questions

We examined reliability from previously conducted job analyses by using the four aforementioned classifications. To explore this purpose, we used meta-analytic procedures. The purpose of these meta-analyses was to estimate the average reliability that one could expect when gathering work information through a job analysis at different data specificities from different sources and when using various descriptive scales. In short, we sought to investigate the following questions: What are the mean estimates of reliability for job analysis information, and how do these estimates differ in magnitude across data specificity, data source, and descriptive scale? Are the levels of interrater reliability higher or lower than levels of intrarater reliability? Finally, does the source of the job analysis information or the choice of descriptive scale affect the magnitudes of reliability estimates?

Method

Database

We conducted a literature search using standard and supplementary techniques in an attempt to lessen the effect of the "file drawer" problem—the increased probability of positive findings in published literature (Rosenthal, 1979). In the case of job analysis research, this could result in unrealistically high estimations of reliability. In addition, many empirical studies about or using job analysis data only report reliability estimations as side bars to the main topic, thus making it more difficult to locate these sources of reliability data. Using the standard technique, we used the Internet and other computer-based resources. Some examples of these sources were PsycINFO, PsychLit, job analysis–related Web sites and listserves, the National Technical Information Services database, as well as other online and offline library databases. Within these sources, we used keyword searches with terms such as "job analysis, job analysis accuracy, job analysis reliability, work analysis, and job information accuracy." The majority of reliability data that we used in this study were gathered with this method. The supplementary technique, meant to expand the breadth of the literature search, used both ancestry and descendency approaches (Cooper, 1984), as well as correspondence with researchers in the field of job analysis. The supplementary approach produced a substantial amount of reliability data in the form of technical reports and unpublished job analyses. Table 1 displays descriptive statistics of the included studies.

Table 1
Descriptive Summary of Collected Data

Data category	No. of studies	No. of reliability estimates
Interrater reliability	31	214
Specificity		
Task	16	119
GWA	15	95
Source		
Incumbent	16	100
Analyst		10
Technical expert		9
Scale		
Frequency	8	10
Importance	10	11
Difficulty	3	10
Time spent	6	23
PAQ	8	83
Intrarater reliability		
Specificity		
Task	10	49
GWA	5	36
Source		
Incumbent	12	42
Analyst	4	31
Technical expert	2	6
Scale		
Frequency	5	13
Importance	4	4
Difficulty	5	7
Time spent	6	10
Publication type		
Journal	26	205
Technical report	10	87
Book	1	3
Dissertation	2	4

Note. GWA = generalized work activity; PAQ = Position Analysis Questionnaire.

Analyses

To be included in the meta-analyses, studies were first required to describe the approach used to assess reliability of the job data. Those that did not assess reliability according to the aforementioned estimation types were excluded. Second, the sample size used in the reliability estimation was required. Third, studies were required to assess the requirements of the job itself, not merely attributes of the workers.

Once the pool of studies was assembled, we coded the data for the purposes of a comparative analysis. Coding allowed for us to conduct separate meta-analyses within each of the study's classifications, hence making the average correlation generated within each grouping more empirically justified. Two raters independently coded the gathered studies according to the four aforementioned classifications. Interrater agreement of study coding was 98%. Disagreements were resolved through discussion, and no additional exclusions were necessary.

We conducted a meta-analysis correcting only for sampling error for each of the distributions gleaned from the study's classifications. When cumulating reliability across several past empirical studies, it may be necessary to determine whether a need to adjust results from various studies to a common length of items or number of raters (interrater reliability) or to a common time interval (intrarater reliability) is required. Two available options were to use the Spearman-Brown formula to bring all estimates to a common length or to use previous research investigating the functional relationship between time intervals and job analysis reliability. The present study conducted meta-analyses both with and without the Spearman-Brown corrections of individual reliability estimates. Without evidence of the functional relationship affecting intrarater reliability of job analysis data, the only statement able to be proffered is that as the time interval increases, reliability generally decreases (Viswesvaran et al., 1996). Thus, no meta-analytic corrections were made to bring estimates of

intrarater reliability to a common time interval. However, intrarater reliability in job analysis can be derived from either a rate–rerate or a repeated item approach. Therefore, to display the potential effects of time and allow comparison between these two common forms of intrarater reliability, separate meta-analyses for repeated item and rate–rerate reliabilities were conducted. The mean time interval for rate–rerate data from the gathered studies was 6.5 weeks and had a range of 1–16 weeks. Rate–rerate data comprised 84% of the collected intrarater reliability data and repeated item data made up the remaining 16%.

As for a common length of items or number of raters, the body of literature on job analysis procedures does not concede a particular recommendation. Suggestions for item length and number of raters varies depending on the organization, project purposes, and the practical limitations of the project (Levine & Cunningham, 1999). Therefore, to portray the potential magnitude change in job analysis reliability as the number of raters fluctuates, we used the Spearman-Brown formula to bring estimates of reliability to several equal numbers of raters (e.g., 5, 15, and 25 raters). As for a common length of items, the Spearman-Brown was similarly used to bring the number of items to several common item lengths (e.g., 100, 200, and 300 items). Because of the smaller number of items typically duplicated in the repeated item approach as opposed to the rate–rerate approach to intrarater reliability (i.e., small subset of items vs. an entire instrument), estimates for these meta-analyses were corrected to the same equal numbered rater sets, but unlike the previous meta-analyses the correction for item length was to 25 items only. The rationale for designating each of these particular rater and item sets was to mirror specifications typically found in job analysis projects. However, we do recognize that these numbered sets are somewhat arbitrary, and others are clearly possible.

For any meta-analysis using reliability estimates corrected with the Spearman-Brown formula, all corrections were applied to the individual reliability estimates. Operationally, individual reliability estimates were first corrected to bring the estimates to equal numbers of raters. Once the individual reliability estimates were corrected for number of raters, they were then corrected for number of items. The individual reliability estimates derived from these various corrections were then used as input for ensuing meta-analyses.

Similar to past research cumulating reliability estimates (Viswesvaran et al., 1996), at least four estimates in a given distribution were needed to perform a meta-analysis. For each meta-analysis conducted, we computed the sample-size weighted mean, observed standard deviation, and residual standard deviation. We also computed the unweighted mean and standard deviation, which do not account for the sample sizes of included estimates. Because each reliability coefficient was weighted, the sample-size weighted mean provided the best estimate of the average reliability for a given distribution, whereas the unweighted mean ensured that the results were not skewed by a few large sample estimates. It is important to note that as a general definition, an intrarater reliability estimate is computed as a sample size of one, and thus sample-size weighted mean intrarater reliability may seem incorrect. However, all of the collected intrarater reliability data were in the form of averages of multiple single-rater reliabilities. Therefore, for intrarater reliability, the sample size of a given averaged intrarater reliability estimate served as the meta-analytic sample-size weight.

Using the results from the statistics described above, we assessed the sampling error variance associated with the mean of the reliability by dividing the variance by the number of estimates averaged (Callender & Osburn, 1988). An 80% confidence interval was calculated from the sampling error of the mean around each mean reliability estimate. We also computed 80% credibility intervals for both interrater and intrarater reliabilities. We calculated these intervals using the sampling error of the mean correlation as derived from the residual standard deviation. We calculated the residual standard deviation as the square root of the difference between observed and sampling error variance. Using the residual standard deviation and the mean reliability correlation to form the 80% credibility interval, we estimated the reliability below which the population reliability value is likely to fall with the chance of .90. The credibility interval refers to the estimated distribution of the population values, not observed values, which are affected by sampling error (Hunter & Schmidt, 1990).

Results

Interrater Reliability

The results for the interrater reliability meta-analyses are reported in the left half of Table 2. The sample size weighted mean reliability estimate for task-level job analysis data was .77 (n = 24,656; k = 119). The sample-size weighted mean reliability estimate for GWA-level job analysis data was .61 (n = 9,999; k = 95). These mean interrater reliability estimates can be seen as the average values one could expect when collecting job analysis information at the respective data specificity. Also shown are the unweighted mean estimates, standard deviations, and the 80% confidence and credibility intervals. Table 2 also provides the results of meta-analyses for interrater reliability classified by source and descriptive scale nested within data specificity. As can be seen, there were insufficient data to perform meta-analyses for GWA data from technical experts and on the scales of importance, difficulty, and time spent. Note that results in Table 2 are not corrected for item length or number of raters.

Table 3 displays the sample-size weighted mean interrater reliabilities corrected to an equal number of raters and items using the Spearman-Brown formula. Similar to the uncorrected estimates, tasks generally have higher interrater reliability than do GWAs. However, for smaller numbers of raters and items (i.e., 5 raters at 100 and 200 items) the interrater reliability for GWA data is slightly higher than for task data. As for data source, analysts tend to show the highest interrater reliability, and incumbents the lowest, regardless of data specificity. Both incumbents and analysts did display higher interrater reliability for tasks than GWAs, although the estimates for larger numbers of incumbents and items were quite comparable across data specificity (e.g., .74 for tasks vs. .73 for GWAs). For interrater reliability in the category of descriptive scale, only scales of frequency had sufficient data to allow comparison across specificity. Here, frequency ratings of GWAs had higher interrater reliability than ratings for tasks. These results should be interpreted with caution, however, due to the small number of reliability estimates (k = 4). Specifically with task data, scales of importance showed the highest levels of interrater reliability. Interestingly, data from scales of difficulty were not the lowest in reliability magnitudes as with the uncorrected estimates. Taken collectively, the evidence from Table 3 generally supports prior suggestions and findings of differential interrater reliability of job analysis data across data specificity, data source, and descriptive scale.

Intrarater Reliability

The right half of Table 2 displays the results of the meta-analyses conducted for intrarater reliabilities of job analysis data. The sample-size weighted mean reliability estimate for task-level job analysis data was .68 (n = 7,392; k = 49). The sample-size weighted mean reliability estimate for GWA-level job analysis data was .73 (n = 3,096; k = 36). Again, these mean intrarater

Table 2
Meta-Analyses of Job Analysis Interrater and Intrarater Reliabilities

	Interrater reliabilities									Intrarater reliabilities								
Data category	*n*	*k*	R_{wt}	SD_{wt}	R_{unwt}	SD_{unwt}	80% CI	SD_{res}	80% CD	*n*	*k*	R_{wt}	SD_{wt}	R_{unwt}	SD_{unwt}	80% CI	SD_{res}	80% CD
Task	24,656	119	.771	.249	.695	.265	.741, .800	.247	.452, 1.000	7,392	49	.684	.107	.723	.131	.664, .703	.097	.557, .810
Source																		
Incumbent	24,420	100	.773	.249	.717	.266	.741, .805	.247	.453, 1.000	7,281	42	.683	.107	.715	.132	.662, .704	.099	.556, .810
Analyst	162	10	.631	.142	.663	.219	.573, .689	.061	.552, .709									
Technical expert	77	9	.466	.178	.479	.195	.390, .543	.222	.179, .754	57	6	.813	.096	.800	.104	.762, .863	.066	.727, .898
Scale																		
Frequency	1,311	10	.699	.159	.588	.224	.634, .764	.153	.501, .896	1,171	8	.721	.034	.771	.086	.706, .736	.021	.694, .748
Importance	280	11	.771	.258	.615	.355	.670, .871	.250	.448, 1.000	1,351	4	.730	.027	.772	.048	.712, .747	.008	.719, .740
Difficulty	746	10	.632	.140	.634	.246	.575, .690	.121	.476, .789	642	7	.485	.075	.579	.134	.448, .521	.029	.448, .522
Time spent	10,785	23	.665	.281	.708	.249	.590, .741	.279	.305, 1.000	1,743	10	.645	.101	.718	.120	.604, .686	.090	.529, .762
GWA	9,999	95	.606	.217	.669	.202	.577, .635	.208	.338, .874	3,096	36	.733	.135	.711	.141	.703, .762	.125	.569, .896
Source																		
Incumbent	1,126	7	.745	.056	.752	.094	.718, .772	.043	.689, .801									
Analyst	8,770	86	.587	.224	.660	.207	.556, .618	.214	.311, .863	2,416	31	.705	.130	.697	.140	.675, .735	.116	.555, .855
Technical expert																		
Scale																		
Frequency	400	4	.688	.031	.808	.110	.668, .708	.043	.633, .743									
Importance																		
Difficulty																		
Time spent																		
PAQ	8,768	83	.587	.223	.657	.207	.555, .618	.213	.311, .862	1,863	27	.679	.137	.819	.092	.645, .713	.120	.524, .834

Note. *k* = number of reliabilities included in meta-analysis; *R* = mean reliability; wt = sample-size weighted; unwt = sample-size unweighted; CI = confidence interval; res = residual; CD = credibility interval; GWA = generalized work activity; PAQ = Position Analysis Questionnaire.

Table 3
Meta-Analyses of Interrater Reliability Estimates Using Spearman-Brown Corrections

	100 items						200 items						300 items					
	5 raters		15 raters		25 raters		5 raters		15 raters		25 raters		5 raters		15 raters		25 raters	
Data category	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*
Task	.312	.253	.488	.294	.570	.295	.421	.285	.598	.293	.674	.283	.487	.294	.659	.286	.727	.279
Source																		
Incumbent	.388	.319	.541	.331	.610	.326	.484	.330	.634	.324	.695	.315	.541	.331	.683	.317	.738	.306
Analyst	.763	.221	.871	.126	.914	.088	.827	.161	.926	.077	.952	.051	.871	.126	.941	.055	.967	.035
Technical expert	.488	.175	.717	.136	.802	.105	.638	.157	.828	.094	.886	.066	.717	.136	.876	.071	.920	.047
Scale																		
Frequency	.290	.312	.423	.335	.492	.337	.371	.329	.517	.335	.587	.324	.423	.335	.573	.328	.642	.308
Importance	.433	.245	.627	.272	.702	.273	.559	.267	.726	.273	.782	.273	.627	.272	.772	.273	.817	.273
Difficulty	.370	.225	.574	.260	.659	.263	.500	.252	.686	.263	.754	.260	.574	.260	.741	.261	.797	.256
Time spent	.227	.238	.373	.301	.446	.326	.317	.281	.471	.334	.538	.354	.373	.303	.525	.350	.586	.366
GWA	.384	.409	.472	.394	.524	.379	.436	.403	.543	.373	.600	.356	.472	.394	.588	.360	.646	.342
Source																		
Incumbent	.286	.230	.439	.291	.527	.263	.377	.301	.560	.249	.655	.205	.439	.291	.635	.215	.727	.167
Analyst	.393	.419	.474	.402	.522	.388	.441	.411	.540	.383	.592	.367	.474	.402	.583	.370	.637	.352
Scale																		
Frequency	.594	.313	.723	.323	.769	.301	.681	.328	.785	.290	.827	.248	.723	.323	.819	.258	.860	.209
PAQ	.365	.417	.442	.398	.489	.384	.409	.408	.507	.379	.554	.364	.442	.398	.551	.367	.607	.351

Note. Only categories with sufficient data to perform a meta-analysis are shown. *R* = mean reliability; GWA = generalized work activity; PAQ = Position Analysis Questionnaire.

reliability estimates can be viewed as the average values one could expect when collecting job analysis information. As evident in Table 2, there were insufficient data to conduct a meta-analysis of task data from analysts. Moreover, only the analyst and PAQ categories had sufficient GWA-level data to allow meta-analysis.

Shown in Table 4 are the sample-size weighted mean intrarater reliabilities corrected for the number of raters and items using the Spearman-Brown formula. Unlike the uncorrected intrarater reliability estimates, when correcting for number of items and raters, task data had higher intrarater reliability than GWA data. No cross-specificity comparisons could be made for the source category due to insufficient data. For tasks only, technical experts displayed higher levels of intrarater reliability than did incumbents. One note of caution, however, is that the sample size for technical experts was rather small ($n = 57$, $k = 6$). Similar to interrater reliability, scales of frequency produced the highest estimates of intrarater reliability for task data but were comparable to scales of importance when the item and rater sets were large. The data displayed in Table 4 suggest that using tasks versus GWAs may increase the intrarater reliability of job analysis data. Insufficient data were available to provide any evidence regarding differences due to source or descriptive scale across data specificity.

Table 5 displays the results of the two additional meta-analyses conducted to compare intrarater reliability estimates derived from repeated item data and estimates from rate–rerate data. The sample-size weighted mean estimate for rate–rerate data was .69 ($n = 7{,}520$; $k = 71$) and .72 ($n = 2{,}968$; $k = 14$) for repeated item data with 80% confidence intervals of .685–.691 and of .721–.723, respectively. When these estimates were corrected to equal numbers of raters and items, the mean reliabilities for repeated item data were much higher than for the rate–rerate data. It should be noted, however, that the magnitude of the discrepancy between reliabilities might be an artifact of the Spearman-Brown formula used for correction. As rate–rerate data come from readministration of entire job analysis instruments, reliability estimation will be based on a greater number of items than the smaller subsets of items used in the repeated item approach. Thus, the Spearman-Brown correction will be much greater for the rate–rerate data than for the repeated item data. In addition, one would expect more intrarater reliability in ratings occurring at the same administration than for those occurring at a second administration.

Comparing the Types of Reliability Estimates

From the values given in Tables 3 and 4, intrarater reliabilities for task data were higher than their interrater reliability counterparts. This suggests that ratings of more specific data may exhibit higher levels of stability than they will levels of consistency. Thus, rating tasks may foster information that remains stable for raters more so than fostering a common rating consensus across raters. Contrarily, ratings of GWA data had higher interrater reliabilities than intrarater estimates. This evidence suggests that at the more general level of activity, ratings are more consistent than they are stable among raters. Hence, ratings of GWAs appear to promote consensus more so than stability. When reviewing the reliabilities across data sources, incumbents seem to provide ratings that are similar in terms of interrater and intrarater reliability levels, whereas analysts display more interrater than intrarater reliability.

Table 4
Meta-Analyses of Intrarater Reliability Estimates Using Spearman-Brown Corrections

Data category	100 items						200 items						300 items					
	5 raters		15 raters		25 raters		5 raters		15 raters		25 raters		5 raters		15 raters		25 raters	
	R	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*	*R*	*SD*
Task	.395	.322	.542	.345	.606	.343	.488	.342	.627	.340	.685	.327	.542	.345	.674	.330	.728	.311
Source																		
Incumbent	.380	.314	.525	.355	.585	.316	.473	.344	.605	.361	.658	.354	.525	.355	.648	.357	.697	.342
Technical expert	.814	.079	.927	.101	.955	.021	.880	.091	.954	.103	.972	.023	.915	.067	.969	.104	.981	.016
Scale																		
Frequency	.512	.347	.637	.370	.681	.374	.595	.363	.694	.374	.728	.369	.637	.370	.722	.371	.752	.358
Importance	.375	.322	.527	.325	.601	.310	.470	.329	.627	.302	.698	.272	.527	.325	.684	.279	.752	.242
Difficulty	.415	.305	.564	.359	.619	.371	.513	.343	.636	.374	.679	.375	.564	.359	.671	.375	.709	.368
Time spent	.350	.330	.472	.384	.522	.394	.429	.370	.539	.395	.585	.391	.472	.384	.576	.393	.621	.382
GWA	.134	.161	.269	.198	.358	.206	.211	.187	.393	.207	.497	.203	.269	.198	.475	.204	.581	.191
Source																		
Analyst	.164	.238	.287	.240	.370	.233	.242	.404	.230	.505	.215	.287	.240	.483	.218	.588	.198	
Scale																		
PAQ	.091	.054	.222	.106	.314	.129	.163	.085	.351	.136	.464	.147	.222	.106	.440	.146	.556	.147

Note. Only categories with sufficient data to perform a meta-analysis are shown. *R* = mean reliability; GWA = generalized work activity; PAQ = Position Analysis Questionnaire.

Table 5
Comparing Rate–Rerate and Repeated Item Reliability Approaches

Reliability approach	Main analysis								Corrected estimates					
									5 raters		15 raters		25 raters	
	n	*k*	R_{wt}	SD_{wt}	R_{unwt}	SD_{unwt}	80% CI	SD_{res}	80% CD	*R*	*SD*	*R*	*SD*	*R*	*SD*
Repeated items	2,968	14	.722	.057	.849	.033	.721, .723	.047	.678, .700	.438	.256	.625	.288	.696	.292
Rate–rerate	7,520	71	.689	.133	.826	.083	.685, .691	.123	.676, .701	.102	.168	.190	.237	.245	.262

Note. The corrected estimates were stepped down to 25 items. *k* = number of reliabilities included in meta-analysis; *R* = mean reliability; wt = sample-size weighted; unwt = unweighted; CI = confidence interval; res = residual; CD = credibility interval.

However, at the same time, incumbents also tended to display the lowest interrater and intrarater reliabilities. Thus, incumbents seem to provide equally consistent and stable ratings, albeit at lower overall reliability levels. As for descriptive scales in relation to task data, scales of frequency and importance generally displayed the highest interrater and intrarater reliability, whereas time-spent scales displayed the lowest interrater and intrarater reliability. In addition, scales of frequency were found to have lower interrater reliability than the importance scales, and both the interrater and intrarater reliabilities of difficulty scales were similar in magnitude to those for scales of importance. It is interesting to note that one salient finding within the task data was that descriptive scales dealing with perceptions of relative value (importance and difficulty scales) tended to have similar and relatively high interrater reliability levels, whereas descriptive scales involving temporal judgments (frequency and time-spent scales) displayed similar and relatively low interrater reliability levels.[1] This evidence suggests that descriptive scales that require respondents to make psychologically similar perceptual judgments tend to foster more agreement across those respondents. However, these similar reliability profiles were not replicated within the intrarater reliability data, thus suggesting the absence of a similar effect for increasing rating stability. Finally, we found ratings from the PAQ to be higher interrater reliability than intrarater reliability.

Discussion

Prior to discussing the implications of this study's results, it is important to note that we are aware of the recent debate surrounding potential inadequacies of the classical measurement model (for discussions of the general disagreements and implications pertaining to the classical measurement model and reliability see K. R. Murphy & De Shon, 2000a; 2000b; Schmidt, Viswesvaran, & Ones, 2000). Although we feel that this is an important debate, an extensive deliberation of these issues is beyond the scope of this article (for a treatment of reliability and validity within the specific context of job analysis, see Morgeson & Campion, 2000; Harvey & Wilson, 2000; Sanchez & Levine, 2000). We clearly used a classical measurement approach in the present study. The rationale for choosing this classical approach coincided with the study's intended purpose, which was to assess average levels of job analysis reliability. The main limitation of using a classical measurement approach lies within the assumptive treatment of all raters or items as equivalent, interchangeable entities. Thus, the results presented herein should be interpreted with caution in that they are inextricably linked to the extent that this assumption is indeed tenable. Nonetheless, the assumption of equivalency does allow the meta-analytic accumulation of reliability data across studies as well as the application of the Spearman-Brown prophecy formula, which both provided potentially useful results for practitioners.

In fact, an interesting question to ask would be whether a classical measurement approach is indeed more appropriate in the context of job analysis. A good argument can be made that a

[1] We thank an anonymous reviewer for this observation.

classical approach may be more amenable to job analysis research than other veins of industrial–organizational psychological research. First, jobs, or any organizational work role for that matter, are institutional conveniences used for purposes suiting the organization itself (i.e., compensation, selection, training, and so forth), and as such, jobs are environmentally defined entities. In other words, jobs generally do have identifiable parameters, at least to the extent to which the responsibilities or desired work outcomes of a given work role are known by the organization. It is rather unlikely to expect that individuals are simply thrust into jobs without any delineation as to the work outcomes expected by the employing company. Second, a job does not necessarily fit the notion of a psychological construct as does job performance (e.g., citizenship and task performance). A psychometric analogy may be that jobs are more akin to principle components than they are to common factors in that a component is entirely defined by its associated composite of items, similar to a job and its tasks or GWAs, whereas a common factor is represented by the communality of various indicators. Third, job analysis research is typically not engaged in model testing of underlying latent variables but rather is focused on the identification of the work activities inclusive of a work role. Of course, these three arguments for the appropriateness of the classical approach in the context of job analysis are likely to become unsubstantiated when research focus moves to more macrounits of analysis in the world of work (e.g., occupations or occupational clusters) or when interest turns to more construct-like entities such as knowledge, skills, and abilities within the general purview of job specification research.

Job analysis data provide the foundation for a wide range of human resource system functions. The results of this study's reliability analyses also provide a general assessment of an important component of job analysis data quality. It is important to note that these analyses should prove beneficial to practitioners as the various levels of reliability surrounding different data specificity, data sources, descriptive scales, and number of raters are now available to aid in the design of future job analysis projects. For instance, when only a certain amount of financial resources are procurable to conduct a job analysis, one could use the information presented herein to provide an estimate of how much reliability could be expected from using 25 incumbents versus 5 trained analysts rating tasks as opposed to GWAs.

One possible reaction to the differing reliabilities found across various data sources may be to infer that it is not unreasonable to conclude that raters do in fact differ, and these differences could also be viewed as not necessarily surprising. However, this postulation raises the issue of how and who defines what constitutes a *job*. It has been our experience that employers often express considerable alarm when informed of unreliable responses within a job analysis for the same job. Most employers certainly hold the expectation that incumbents—for many kinds of work—should be seeing the same workplace behaviors. The expectation of a common consensus surrounding work behavior is especially true of the more specific task data. Examining the interrater reliability levels found herein speaks to the issue of a common consensus. Because reliability levels were generally higher for tasks than GWAs, there appears to be a greater level of agreement in ratings of tasks as opposed to GWAs. This would seem rational in that tasks, as more specific pieces of work data, are most likely easier to interpret and recognize as being performed within an individual's job than less specific GWAs.

An additional implication that arises from the interrater reliability difference between tasks and GWAs falls particularly within the context of the contemporary and prevalent shift to using more abstract descriptors when analyzing work, such as in competency modeling and O*NET (Jeanneret, Borman, Kubisiak, & Hanson, 1999). From our results, it seems that the use of more abstract data could lead to a decline in agreement across respondents. Agreement is particularly salient in terms of job analysis data. In fact, agreement between raters in job analysis is perhaps more of a concern than with job-performance data, such as in multi-source performance assessment in which low agreement may be attributed to different perspectives on performance and may be expected and even desirable. Unlike job-performance data, job analysis data serve as an initial informational input into a given human resource system. To the extent that other human resource functions are based on the underpinnings and delineations of collected job analysis data, the more desirable greater agreement (i.e., higher interrater reliability) should be on the various job tasks or duties that make up an organization's defined work roles. Moreover, the potential decline in agreement in relation to abstraction of data specificity may also serve to exacerbate issues surrounding low levels of interrater reliability as representing or masking true differences in job composition (see Harvey, 1991). An important question that becomes apparent and needs further investigation pertains to the actual cost, such as practicality or legality, associated with the loss of interrater reliability when one chooses to use more abstract levels of analysis. Of course these costs must be considered in the direct context of the intended use of the job analysis information (i.e., designing training, writing job descriptions, and so forth).

Our interpretation of the reliability findings should not be seen as necessarily advocating a move toward task-oriented job analyses, rather we believe the true value of job analysis depends directly on the purpose for which the data are collected. If one is designing a training program or developmental feedback, task data are indeed essential. Any practitioner who has conducted task analysis is well aware that it is particularly time consuming, expensive, and to be avoided whenever possible. We do feel job analysis data have several potential uses in which broader and more strategic views of work may be sufficient, such as in criterion development. However, for some job analysis purposes, a detailed examination of tasks is unavoidable. Considering the reliability results within this study, incorporating at least some task information into more general analyses may prove to bolster reliability. It is interesting to note that this may already be happening in practice, although the intention is not explicitly stated, in that currently there is a concerted effort to link task data to more generic O*NET descriptors.

Issues surrounding the stability of job analysis ratings are also relevant from the estimates of intrarater reliability, in which the reliabilities for ratings of tasks are more stable than ratings for GWAs. These results are congruent to recent suggestions that, as one moves from the molecular to the molar in describing work, the reliability decreases (Harvey & Wilson, 2000; K. F. Murphy & Wilson, 1997). However, this hypothesis cannot be completely confirmed due to the nature of the cumulated data. A clear discrepancy between estimates derived from the two approaches to

assessing intrarater reliability can be seen in Table 5, with the repeated item technique resulting in higher estimates. The nature of the data is such that the vast majority of the repeated item data are tasks, whereas the rate–rerate data are blended with tasks and GWAs. Hence, the higher reliability associated with task data could be more a factor of the employed intrarater reliability assessment technique than of the specificity of the data. The differences due solely to the applied technique notwithstanding, we derived a large portion of intrarater reliability data at the GWA-level from analysts (e.g., 31 of the 36 studies). One can argue that trained job analysts could be expected to possess a more stable schema of various types of abstract descriptors, in this case GWAs, stemming from their experience in analyzing a wider range of jobs as compared with their incumbent counterparts. Moreover, of these 31 studies, 27 used the PAQ, which is one of the most well researched and developed generic job analysis instruments available and likewise could be expected to foster more stable ratings. Finally, considering the relatively short average time interval of the rate–rerate data (e.g., 6.5 weeks) the possibility of changing activities in the make-up of jobs within the cumulated studies is largely eliminated. The general conclusion when considering these various extraneous influences regarding the intrarater reliability found within this study is that additional research is clearly needed if a more accurate picture is to be drawn.

A supplemental finding was revealed through conducting the literature search, and it is important to emphasize. There were at least three studies found in which actual availability of data would have allowed for an easy assessment of reliability, yet no such analyses were performed or reported. Moreover, studies were also found that claimed to investigate the quality of job analysis data, yet gave no mention of reliability within their manuscripts. These findings were perplexing to say the least, considering the long-standing tradition of reporting reliability for any psychometric device, especially when it is used for decision-making purposes, not to mention the potential legal ramifications. Furthermore, we recognize that although exhaustive efforts were made to capture the maximum amount of available job analysis reliability information, we were only able to locate 46 studies (299 individual reliability estimates) providing usable reliability data. This shortage of reliability data for purposes of meta-analysis is also one of the greatest limitations of the present study.[2]

Unlike much previous job analysis research, the present study did not analyze any specific individual characteristics that may affect reliability. However, magnitude differences were indeed evident across the various data sources for both interrater and intrarater reliabilities of tasks and GWAs. More research is needed to investigate potential individual characteristics affecting job analysis reliability. One promising avenue not included within this study's examinations is that of organizational tenure or amount of job experience (Tross & Maurer, 2000). Recent research has suggested that the quality of job analysis data may have the classic inverted-U relationship (Dierdorff, Carter, & Wilson, 2001; K. F. Murphy, 2000), with more reliable data, using the repeated item approach, coming from individuals within the medium range of tenure versus those with extremely short or long lengths of tenure (e.g. rookies and seasoned veterans). This avenue of research appears propitious for future investigations.

Some of the specific findings of this study also help point to many potentially fruitful avenues for future job analysis research. These avenues are realized through the empirical gaps encountered within the data collection, particularly those gaps associated with availability of reliability estimates. The shortage of reliability data in several areas was evident (see Table 2). For instance, more interrater reliability data are needed at the GWA-level from technical experts, as well as more intrarater reliability data at the GWA-level from both technical experts and incumbents. These areas would seem highly salient as they are at the more generic level of data and use sources that are more commonly chosen in contemporary job analysis projects. The quantity of intrarater reliability estimates are also lacking for task ratings from analysts and technical experts.

The present study incorporated reliability information that was derived from a myriad of jobs (e.g., aircraft mechanics, police officers, insurance agents, and so forth). Indeed, jobs may vary in terms of complexity, and this complexity could possibly affect job analysis reliability. The present study made no attempt to code for level of job complexity. Research investigating the influences of job complexity on job analysis reliability would be useful along with the impact of shift work and disparate locations. In addition, future meta-analytic research focusing on the reliability of worker-oriented data, as used in job specifications (e.g., KSAs), would also be an important endeavor. The GWA data used in the present study were more descriptively concrete than abstract descriptors such as KSAs, which have been shown to result in lower reliability estimates (K. F. Murphy & Wilson, 1997). Additional examination of these issues is warranted.

Gael (1983) pointed out that the computing of reliability of job analysis data is an important, but not a sufficient, condition for displaying data validity. Reliability becomes a precondition for validity and is necessary for developing any valid prediction process on the basis of job analysis results, an important point when considering the many human resource functions based on the results of job analyses. Unfortunately, the present study could not directly incorporate meta-analytic investigations into issues of validity, mainly because of the lack of sufficient data. There are at least two explanations for the lack of validity studies for job analysis data. One is that they have not been done and hence cannot be reported. The other is that they have been done and the results are not good so they reside in a file drawer, unreported in the hope that it is an artifact. Given the foundational nature of job analysis data in industrial–organizational psychology, it is important to ascertain which explanation is the case. Moreover, what work has been done is limited but not particularly encouraging (Wilson, 1997). Future research is clearly needed to further investigate more functional methods of assessing validity of job analysis data, as well as developing a theoretical model of the determinants of job analysis quality.

In conclusion, we hope that the results given by this study will provide reference points for both researchers and practitioners of job analysis on which to compare obtained reliability levels. Future research should continue to investigate ways in which reliability levels can be increased, as well as empirically examine those areas

[2] We would greatly appreciate any additional reliability data that were not included within our meta-analyses but that are derived from job analysis and fit the classifications described within this study. Please contact Mark A. Wilson for further information.

lacking sufficient data to be analyzed by the present study. In addition, we hoped that future users of job analysis data will more frequently take the time to conduct reliability analyses. Research focusing on practical ways of assessing the validity of collected job analysis data is also needed. The importance of examining the quality of collected job analysis data through reliability and validity cannot be overstated.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Aamodt, M. G., Kimbrough, W. W., Keller, R. J., & Crawford, K. J. (1982). Relationship between sex, race, and job performance and the generation of critical incidents. *Education and Psychological Research, 2,* 227–234.

*Ash, R. A. (1982). Job elements for task clusters: Arguments for using multi-methodological approaches to job analysis and a demonstration of their utility. *Public Personnel Management, 11,* 80–89.

*Banks, M. H., & Miller, R. L. (1984). Reliability and convergent validity of the Job Components Inventory. *Journal of Occupational Psychology, 57,* 181–184.

*Birt, J. A. (1968). The effect of the consistency of job inventory information upon simulated airmen reassignment. *Dissertation Abstracts International, 29,* 3121B. (UMI No. 69–2890)

Borman, W. C., Dorsey, D., & Ackerman, L. (1992). Time-spent ratings as time allocation strategies: Relations with sales performance in a stockbroker sample. *Personnel Psychology, 45,* 763–777.

*Butler, S. K., & Harvey, R. J. (1988). A comparison of holistic versus decomposed rating of Position Analysis Questionnaire work dimensions. *Personnel Psychology, 41,* 761–771.

*Cain, P. S., & Green, B. F. (1983). Reliabilities of selected ratings available from the Dictionary of Occupational Titles. *Journal of Applied Psychology, 68,* 155–165.

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

*Chatfield, R. E., & Royle, M. H. (1983). *Methods to improve task inventory construction (PRDC No. TR-83–36).* San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. AD-A144–450)

*Christal, R. E. (1971). *Stability of consolidated job descriptions based on task inventory survey information (AFHRL No. TR-71–48).* Lackland Air Force Base, TX: Personnel Research Laboratory. (NTIS No. AD-734-739)

Cooper, H. (1984). *The integrative research review: A systematic approach.* Beverly Hills, CA: Sage.

*Cornelius, E. T., & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. *Journal of Applied Psychology, 65,* 155–163.

*Cragun, J. R., & McCormick, E. J. (1967). *Job inventory information: Task and scale reliabilities and scale interrelationships (PRL No. TR-76–15).* Lackland Air Force Base, TX: Personnel Research Laboratory. (NTIS No. AD-681–509)

Cunningham, J. W. (1996). Generic job descriptors: A likely direction in occupational analysis. *Military Psychology, 8,* 247–262.

*Cunningham, J. W., Boese, R. R., Neeb, R. W., & Pass, J. J. (1983). Systematically derived work dimensions: Factor analysis of the Occupation Analysis Inventory. *Journal of Applied Psychology, 68,* 232–252.

Cunningham, J. W., Drewes, D. W., & Powell, T. E. (1995). Framework for a revised Standard Occupational Classification (SOC). In Standard Classification Revision Policy Committee (Eds.), *Seminar on research findings* (U.S. Department of Labor No. 1995–398–319/40067, pp. 57–165). Washington, DC: U.S. Government Printing Office.

*Cunningham, J. W., Wimpee, W. E., & Ballentine, R. D. (1990). Some general dimensions of work among U.S. Air Force enlisted occupations. *Military Psychology, 2,* 33–45.

*DeNisi, A. S., Cornelius, E. T., & Blencoe, A. G. (1987). Further investigation of common knowledge effects on job analysis ratings. *Journal of Applied Psychology, 72,* 262–268.

*Dierdorff, E. C., Carter, L., & Wilson, M. A. (2001). *An SBI field agent selection system* (Tech. Rep. No. 2000–0700). Raleigh, NC: North Carolina State University.

Dunnette, M. D. (1976). *Handbook of industrial and organizational psychology.* Chicago, IL: Rand McNally.

*Dunnette, M. D., & Kirchner, W. K. (1959). A checklist for differentiating different kinds of jobs. *Personnel Psychology, 12,* 421–444.

*Friedman, L. (1991). Degree of redundancy between time, importance, and frequency task ratings: Correction. *Journal of Applied Psychology, 76,* 366.

*Friedman, L., & Harvey, R. J. (1986). Can raters with reduced job descriptive information provide accurate Position Analysis Questionnaire (PAQ) ratings? *Personnel Psychology, 39,* 779–789.

*Frieling, E., Kannheiser, W., & Lindberg, R. (1974). Some results with the German form of the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 59,* 741–747.

Gael, S. (1983). *Job analysis: A guide to assessing work activities.* San Francisco: Jossey-Bass.

*Geyer, P. D., Hice, J., Hawk, J., Boese, R., & Brannon, Y. (1989). Reliabilities of ratings available from the Dictionary of Occupational Titles. *Personnel Psychology, 42,* 547–560.

*Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39,* 543–564.

*Green, S. B., & Veres, J. G., III (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology, 5,* 47–61.

Gutman, A. (1993). *EEO law and personnel practices.* Thousand Oaks, CA: Sage.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J., Friedman, L., Hakel, M. D., & Cornelius, E. T. (1988). Dimensionality of the job element inventory, a simplified worker-oriented job analysis questionnaire. *Journal of Applied Psychology, 73,* 639–646.

Harvey, R. J., & Wilson, M. A. (2000). Yes Virginia, there is an objective reality in job analysis. *Journal of Organizational Behavior, 21,* 829–854.

*Hazel, J. T., Madden, J. M., & Christal, R. E. (1964). Agreement between worker–supervisor descriptions of the worker's job. *Journal of Industrial Psychology, 2,* 71–79.

*Henry, M. S., & Morris, S. B. (2000, April). *Incumbent performance level as a predictor of job analysis ratings.* Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69,* 167–173.

*Hughes, G. L., & Prien, E. P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. *Personnel Psychology, 42,* 283–292.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

*Jeanneret, P. R., Borman, W. C., Kubisiak, U. C., & Hanson, M. A. (1999). Generalized work activities. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 105–125). Washington, DC: American Psychological Association.

*Jones, A. P., Main, D. S., Butler, M. C., & Johnson, L. A. (1982). Narrative job descriptions as potential sources of job analysis ratings. *Personnel Psychology, 35,* 813–828.

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. *Personnel Psychology, 44,* 27–50.

Levine, E. L., & Cunningham, J. W. (1999). How O*NET can cut the work in your work analysis (and make it better too). In D. W. Drewes, M. A. Wilson, & J. W. Cunningham (Eds.), *O*NET work analysis field book: A guide for defining the world of work* (pp. 5–36). Raleigh, NC: National Center for O*NET Development.

Levine, E. L., Sistrunk, F., McNutt, K. J., & Gael, S. (1988). Exemplary job analysis systems in selected organizations: A description of processes and outcomes. *Journal of Business and Psychology, 3,* 3–21.

*Ludlow, L. H. (1999). The structure of the Job Responsibilities Scale: A multimethod analysis. *Educational and Psychological Measurement, 59,* 962–975.

*McCormick, E. J. (1960). Effect of amount of job information required on reliability of incumbents' check-list reports. *USAF Wright Air Development Division Technical Note,* 60–142.

*McCormick, E. J., & Ammerman, H. L. (1960). *Development of worker activity checklists for use in occupational analysis (WADD No. TR-60–77).* Lackland Air Force Base, TX: Personnel Research Laboratory. (NTIS No. AD-248–385)

*McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the position analysis questionnaire (PAQ). *Journal of Applied Psychology, 56,* 347–368.

*McCormick, E. J., & Tombrink, K. B. A. (1960). *A comparison of three types of work activity statements in terms of the consistency of job information reported by incumbents (WADD No. TR-60–80).* Lackland Air Force Base, TX: Personnel Research Laboratory. (NTIS No. AD-248–386)

*Meyer, H. H. (1959). Comparison of foremen and general foreman conceptions of the foreman's responsibilities. *Personnel Psychology, 12,* 445–452.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82,* 627–655.

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior, 21,* 819–827.

Morsh, J. E. (1964). Job analysis in the United States Air Force. *Personnel Psychology, 17,* 7–17.

*Murphy, K. F. (2000). *An examination of factors impacting the accuracy, reliability, and validity of job analysis task ratings.* Unpublished doctoral dissertation, North Carolina State University, Raleigh.

Murphy, K. F., & Wilson, M. A. (1997, April). *Estimating the reliability of job analysis ratings: The role of level of abstraction, method of estimation and modality.* Presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Murphy, K. R., & De Shon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873–900.

Murphy, K. R., & De Shon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53,* 913–924.

*Pass, J. J., & Robertson, D. W. (1980). *Methods to evaluate scales and sample size for stable task inventory information* (Tech. Rep. No. 80-28). San Diego, CA: U.S. Navy Personnel Research & Development Center.

*Patrick, J., & Moore, A. K. (1985). Development and reliability of a job analysis technique. *Journal of Occupational Psychology, 58,* 149–158.

*Pine, D. E. (1995). Assessing the validity of job ratings: An empirical study of false reporting in task inventories. *Public Personnel Management, 24,* 451–459.

*Ployhart, R. E., Schmitt, N., & Rogg, K. (2000, April). *Linking job analysis ratings to firm performance: Relations between supervisory experiences, roles, tasks, and firm performance.* Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Richmann, W. L., & Quinones, M. A. (1996). Task frequency rating accuracy: The effect of task engagement and experience. *Journal of Applied Psychology, 81,* 512–524.

Rosenthal, R. (1979). Meta-analytic procedures for social research. Newbury Park, CA: Sage.

*Russell, T. L., Crafts, J. L., Tagliareni, F. A., McCloy, R. A., & Barkley, P. (1994). *Job analysis of special forces jobs (FR-PRD-94–19).* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

*Sanchez, J. I., & Levine, E. L. (1989). Determining important tasks within jobs: A policy-capturing approach. *Journal of Applied Psychology, 74,* 336–342.

*Sanchez, J. I., & Levine, E. L. (1994). The impact of rater's cognition on judgment accuracy: An extension into the job analysis domain. *Journal of Business and Psychology, 9,* 47–57.

Sanchez, J. I., & Levine, E. L. (2000). Accuracy or consequential validity: Which is the better standard for job analysis data? *Journal of Organizational Behavior, 21,* 809–818.

Schippmann, J. S. (1999). *Strategic job modeling: Working at the core of integrated human resources.* Mahwah, NJ: Erlbaum.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53,* 901–912.

*Schmitt, N., & Cohen, S. A. (1989). Internal analyses of task ratings by job incumbents. *Journal of Applied Psychology, 74,* 96–104.

*Schmitt, N., & Fine, S. A. (1983). Inter-rater reliability of judgments of functional levels and skill requirements of jobs based on written task statements. *Journal of Occupational Psychology, 56,* 121–127.

Shrout, P. E, & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

*Smith, J. E., & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology, 32,* 677–692.

*Taylor, L. R. (1978). Empirically derived job families as a foundation for the study of validity generalization: Study I. The construction of job families based on the component and overall dimensions of the PAQ. *Personnel Psychology, 31,* 325–340.

*Taylor, L. R., & Colbert, G. A. (1978). Empirically derived job families as a foundation for the study of validity generalization: Study II. The construction of job families based on company-specific PAQ job dimensions. *Personnel Psychology, 31,* 341–353.

*Terry, D. R. (1973). *Methodological study for determining the task content of dental auxiliary education programs* (HRP-0004628). Bethesda, MD: Bureau of Health Manpower Education.

Tross, S. A., & Maurer, T. J. (2000). The relationship between SME job experience and job analysis ratings: Findings with and without statistical control. *Journal of Business and Psychology, 15,* 97–110.

Uniform Guidelines on Employee Selection Procedures, 43 C.F.R. § 38295 (1978).

Viswesvaran, C., Ones, D., & Schmidt, F. L. (1996). A comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Werner, J. M., & Bolino, M. C. (1997). Explaining U.S. courts of appeals decisions involving performance appraisal: Accuracy, fairness, and validation. *Personnel Psychology, 50,* 1–24.

Wilson, M. A. (1997). The validity of task coverage ratings by incumbents and supervisors: Bad news. *Journal of Business and Psychology, 12,* 85–95.

*Wilson, M. A., Harvey, R. J., & Macy, B. (1990). Repeating items to estimate the test–retest reliability of task inventory ratings. *Journal of Applied Psychology, 75,* 158–163.

Received March 27, 2002
Revision received November 14, 2002
Accepted November 20, 2002 ■