Journal of Occupational and Organizational Psychology (2019), 92, 764–786

The British Psychological Society

www.wileyonlinelibrary.com

Job analysis ratings and criterion-related validity: Are they related and can validity be used as a measure of accuracy?

Jeff A. Weekley[1]*, Jeffrey R. Labrador[2], Michael A. Campion[3] and Kathleen Frye[4]

[1]University of Texas at Dallas, Richardson, Texas, USA
[2]IBM, San Diego, California, USA
[3]Purdue University, West Lafayette, Indiana, USA
[4]Independent Consultant, McKinney, Texas, USA

Job analysis data are largely judgements from subject matter experts (SMEs), judgements with unknown accuracy. To date, accuracy has been inferred largely based on inter-rater reliability or agreement between SMEs and without reference to an external criterion. The current research examined job analysis rating accuracy by comparing SME importance ratings of knowledge, skills, abilities, and other requirements (KSAOs) with the validity of measures of these same KSAOs in predicting job performance. We tested hypotheses about whether SME judgement accuracy is moderated by SME job tenure, industry experience, role, self-reported knowledge of the job, and data scrubbing. Four independent tests involving 48 separate validation studies were conducted. In three of the four samples, there was a large (r = .50 range) relationship between trait importance and trait validity, showing that job analysis ratings can be directly related to test validities and serve as a measure of job analysis accuracy. Moderator analyses showed that the best results may come from supervisors, rather than incumbents, and those who know the job extremely well (there were no differences due to SME job tenure, industry experience, or deletion of outliers). Showing a direct relationship between SME judgements and actual criterion-related validity provides a new lens for operationalizing accuracy in job analysis research.

Practitioner Points

- This study demonstrates that test validities can serve as a measure of accuracy, providing a new avenue for job analysis research.
- The most accurate job analysis ratings came from supervisors and those who reported knowing the job extremely well.

Although job analysis underlies many of the initiatives undertaken by organizations to improve individual performance (e.g., selection, performance appraisal, and training), its focus in published research has declined in popularity (Cascio & Aguinis, 2008; Morgeson & Dierdorf, 2011; Sanchez & Levine, 2012). One reason why research on job analysis may

*Correspondence should be addressed to Jeff A. Weekley, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080-3021, USA (email: jeff.weekley@utdallas.edu).

DOI:10.1111/joop.12272

have stalled is that there is no conclusive way of determining 'best practices'. Job analysis data are largely judgements from subject matter experts (SMEs). These may be judgements about the job (e.g., how frequently a task is performed, how difficult that task is to learn, or how big are the consequences associated with performing it incorrectly?) or about the worker (e.g., how important is a particular knowledge, skill, ability, or other characteristic (KSAO) to effective performance and is that KSAO needed at time of hire?). Either way, SMEs are typically rating work activities or worker characteristics on closed-ended scales and these data used to make decisions impacting programmes and people. Morgeson and Campion (1997) identified a host of social and cognitive sources of potential error in job analysis data. One consequence has been an important discussion about the concept of accuracy in job analysis data (Morgeson & Campion, 2000; Sanchez & Levine, 2000). As Sanchez and Levine (2000) noted, there is no 'gold standard' or unambiguously correct depiction of a job's tasks, requirements, or conditions.

The purpose of this research was to begin to fill this void. We report research linking job analysis ratings of KSAOs to the resultant validities of measures of these same KSAOs in predicting job performance. In doing so, we offer a new and potentially better way of operationalizing the concept of 'accuracy' in job analysis conducted for selection purposes – which job analysis practices result in importance ratings most strongly related to criterion-related validities? While much research has been conducted on job analysis, without an accepted criterion of accuracy it has been difficult to determine which methods, sources, or practices provide the best information. Based on prior theory and research, we consider six variables that might moderate the accuracy of job analysis ratings: SME role (incumbent vs. supervisor), SME job tenure, SME industry experience, frequency of SME contact with the job, self-reported SME knowledge of the job, and data scrubbing for outliers. Using the job analysis–validity relationship as our criterion, we illustrate how guidance can be provided as to which are important considerations and which are not.

Accuracy in job analysis ratings

Past research on accuracy, reliability, and agreement

To date, the accuracy of judgements from SMEs, be they judgements about the job or about the worker, has been largely determined without reference to an external criterion. The most common practice has been to rely on measures of consistency, usually either inter-rater reliability or agreement (Morgeson & Campion, 1997). There have been many studies of the reliability of job analysis ratings. For example, in their meta-analysis of 299 reliability coefficients from 49 studies, Dierdorff and Wilson (2003) found inter- and intra-rater reliabilities to be higher for tasks than work activities and that incumbents provided less reliable ratings than job analysts or experts (which included supervisors). Research also suggests that job analysis ratings of worker attributes are reliable, albeit to a lesser extent (e.g., Dierdorff & Morgeson, 2009; Peterson *et al.*, 2001a; Tsacoumis & Van Iddekinge, 2006).

Many studies have suggested that inter-rater differences in work activity ratings may not reflect error, but rather real differences in how incumbents go about doing their jobs (e.g., Borman, Dorsey, & Ackerman, 1992; Dierdorf & Morgeson, 2007; Sanchez, Prager, Wilson, & Viswesvaran, 1998). Similarly, differences in intra-rater reliability may reflect changes over time in how the incumbent approaches the job and not measurement error (Befort & Hattrup, 2003; Borman *et al.*, 1992; Landy & Vasey, 1991; Prien, Prien, &

Wooten, 2003; Sanchez, 1990). There have been several theoretical explanations offered for the reasons for differences in job analysis results. The lack of agreement between SMEs, rather than indicating inaccuracy, may reflect 'job crafting' or real differences in how incumbents mould the job to enhance their outcomes (Bruning & Campion, 2018; Wrzesniewski & Dutton, 2001). Dierdorf and Morgeson (2007) used role theory to explain how interdependence, autonomy, and routinization in the work social context allowed differences in idiosyncratic role enactment. Borman *et al.* (1992) used career development theory to explain learning that occurs over one's career. Finally, Befort and Hattrup (2003) used leadership theory to explain why more experienced managers might emphasize contextual behaviours.

Past research on measuring job analysis accuracy as outcomes

As alternatives to reliability-based measures of accuracy, it has been suggested that researchers shift focus to the consequences associated with the job analysis (Sanchez & Levine, 2000) or the validity of inferences drawn from the job analysis (Morgeson & Campion, 2000). Despite having been suggested more than 18 years ago, relatively little research has been conducted examining the consequences of job analysis (c.f., Jones *et al.*, 2001; Levine, Ash, & Bennett, 1980; Manson, 2004; Van Iddekinge, Raymark, & Eidson, 2011). While research is only beginning to shift from reliability to consequences, examining job analysis methods and results in the light of the purpose they were collected to serve offers an important new lens for research. In a selection context, information is often gathered on the importance of KSAOs for job performance to guide decisions about both the content and method of the selection procedures to be used. This suggests that the accuracy of KSAO importance inferences should be verified by comparison to the accuracy with which these KSAOs predict job performance. When captured for purposes of selection system design, the best measure of job analysis accuracy would be the validity of assessments developed or selected based on the job analysis. The most accurate job analysis information then would be that which is most strongly related to the resulting validities of the KSAOs in predicting performance.

There is surprisingly little research on the relationship between job analysis and validity. Several meta-analyses indicate that assessments based on a job analysis show greater criterion-related validity than those not derived from an analysis of the job or its requirements (e.g., Dye, Reck, & McDaniel, 1993; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Tett, Jackson, & Rothstein, 1991; Weisner & Cronshaw, 1988). However, these studies only examined whether or not a job analysis was conducted. This does not show that job analysis ratings, which are the fundamental data collected in most job analyses, are directly related to criterion-related validity. For example, it could be that the apparent relationship is due to an unmeasured third variable (e.g., competent test developers (1) build valid assessments and (2) conduct job analyses, but would build valid assessments even if they did not conduct the job analyses).

In fact, the few studies that have directly compared job analysis ratings with validities have been far less encouraging. Ostroff and Schmitt (1987) used the criterion-related validities of six dimensions from an in-basket exercise as the dependent variable and the content validity ratio (CVR) of these as the independent variable. Their results showed no relationship between the two, suggesting that the degree to which a construct is essential to the job was unrelated to the validity of that construct in predicting performance. Carrier, Dalessio, and Brown (1990) conducted a similar study using interview ratings collected from three different samples on three different interview forms. Although the

item CVRs were significantly related to item validities in one of the samples, the correlation between the two was negative and not significant in the other two. A number of studies have also looked at the validity of test batteries comprised of cognitive ability measures constructed on the basis of varying degrees of match between test and job content. Peterson, Wise, Arabian, and Hoffman (2001b) found that general composites created from the Armed Services Vocational Aptitude Battery (ASVAB) were usually about as valid as those designed specifically for a job. Other studies of the ASVAB reached similar conclusions (Earles & Ree, 1992; Fairbanks, Welsh, & Sawin, 1990; Jones & Ree, 1998). Consistent with this, Murphy (2009) has argued that when a set of tests are positively correlated with one another and the criterion (a condition called positive manifold), a battery created from any subset of the tests will be valid and often as valid as one derived from an analysis of the content of the job. However, because of their oftentimes small intercorrelations, Murphy (2009; Murphy, Dzieweczynski, & Zhang, 2009) did recognize that content validity might be relevant in identifying personality tests that would be most useful in predicting performance.

Therefore, in the current research we rely on the criterion-related validity of personality measures to first test the fundamental assumption that job analysis ratings from SMEs provide insight into the constructs that should be included in an assessment. To provide a robust test of our hypotheses, we use validities computed in four samples as criteria to evaluate directly the relationship between job analysis ratings and criterion-related validity. Although direct tests of this question (e.g., Carrier *et al.*, 1990; Ostroff & Schmitt, 1987) have suggested otherwise, given (1) the limitations of those studies, (2) the meta-analytic research showing predictors based on a job analysis to be more valid than those not, and (3) the theoretical support for a link between content and criterion-related validities (e.g., Binning & Barrett, 1989), we believe that KSAOs ratings will be related to resulting validities.

Hypothesis 1: Importance ratings of personality constructs will be significantly related to the validities observed when measures of those constructs are correlated with job performance.

Potential moderators of accuracy

Using validity as the measure of accuracy also offers another way of testing theory about the relative merits of different types of SMEs and other job analysis practices. Harvey (1991) called the choice of who to capture job analysis information from as one of the most critical decisions to be made in any job analysis. Specifically, we consider the relative accuracy of five variations in SME composition and one recommended data analytic practice.

SME job tenure and accuracy

The research on SME job tenure has yielded mixed results, with some finding important differences as a function of tenure (Borman *et al.*, 1992; Goldstein, Noonan, & Schneider, 1992; Landy & Vasey, 1991; Richman & Quinones, 1996; Tross & Maurer, 2000), while others finding no difference (Cornelius & Lyness, 1980; Green & Stutzman, 1986; Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989). Most of this research focused on task ratings though, with only two studies comparing KSAO ratings (Goldstein *et al.*, 1992; Tross & Maurer, 2000). Both found a slight tendency for more tenured SMEs to provide higher ratings. Regarding potential accuracy implications, as compared to less tenured SMEs, more tenured SMEs may (1) rate KSAOs differently (Goldstein *et al.*, 1992; Tross &

Maurer, 2000), (2) allocate their time differently (Borman *et al.*, 1992; Landy & Vasey, 1991), and (3) perform better (McDaniel, Schmidt, & Hunter, 1988). Based on this, we expect that tenure should increase knowledge of the job and its requirements and that more tenured SMEs will therefore provide more accurate information on KSAO importance than less tenured ones.

Hypothesis 2: The relationship between importance ratings of personality constructs and observed validities will be stronger when importance ratings are captured from more job-tenured than less job-tenured SMEs.

SME industry experience and accuracy

Research on the effects of experience on job analysis ratings has largely focused on job tenure. Past research has not explicitly examined differences in industry experience, but the research on tenure suggests that SMEs having the benefit of greater industry experience should have a better understanding of context and a broader base of knowledge on what it takes to do the job well. While Dokko, Wilk, and Rothbard (2009) found industry experience to be unrelated to performance, others have consistently found that employers are willing to pay a premium for it (Ang, Slaughter, & Ng, 2002; Goldsmith & Veum, 2002; Parent, 2000). This premium paid for industry experience is presumably for the enhanced knowledge this experience provides, knowledge which should have the same effect on job analysis accuracy as that due to job tenure. Thus, we expect industry experience to be beneficial to SMEs and enable those with high levels of it to provide more accurate ratings than SMEs with less industry experience.

Hypothesis 3: The relationship between importance ratings of personality constructs and observed validities will be stronger when importance ratings are captured from SMEs with more industry experience than from SMEs with less industry experience.

SME role and accuracy

As regards role, job analysis questionnaire information is typically captured from one of three sources: incumbents, supervisors, or job analysts. Dierdorff and Wilson (2003) compared these three sources in their meta-analysis and found that analysts demonstrated the highest inter-rater reliability and incumbents the lowest for both tasks and behaviours. Incumbents also demonstrated the lowest intra-rater reliability. On the other hand, Richman and Quinones (1996) found that incumbents were more accurate than observers (pseudo supervisors) when rating task frequency. Others (Manson, Levine, & Brannick, 2000; Sanchez, 1990) have found there to be little difference between incumbents and supervisors across task inventory ratings, leading them to conclude that the choice of respondent will have little impact on results.

Guder (2012) suggested that incumbents may be better positioned to describe how the work is actually performed, and therefore be better able than supervisors to accurately report the tasks performed on the job. When it comes to job requirements, however, theory and research suggest that supervisors may provide the more accurate information. As opposed to more observable tasks, job requirements often involve more abstract judgements of unobservable psychological constructs such as the importance of ability or personality (Harvey, 1991). Because the accuracy of ability or personality requirements are more difficult to verify, they may be more prone to inflation. Morgeson, Delaney-Klinger, Mayfield, Ferrara, and Campion (2004) confirmed that incumbents' ratings of ability statements were significantly more inflated than were task statement ratings, a

pattern not repeated when ratings were captured from supervisors or job analysts. When an incumbent perceives some benefit from making a positive impression, his/her ratings of job requirements may be inflated and therefore less accurate (Guder, 2012). As Morgeson *et al.* (2004) observed, respondents' ratings may reflect perceptions of their own abilities rather than the abilities required by the job. Supervisors, because they are not rating their own job, should be less prone to this form of self-presentation bias. Further, because supervisors observe the performance and behaviour of multiple incumbents, they may be better suited to identify the attributes that differentiate high from low performers (Guder, 2012). Finally, because supervisors often consider the behaviours and requirements of different jobs, they may benefit from a broader perspective on attribute importance that incumbents do not share. When looking at only one job, such as one's own, every personality trait looks important. For these reasons, we believe that supervisors will more accurately identify the requirements of the job.

Hypothesis 4: The relationship between importance ratings of personality constructs and observed validities will be stronger when importance ratings are captured from supervisors rather than incumbents.

Knowledge of job and accuracy

Subject matter experts may differ in terms of how well they know the job. The rationale for using role, tenure, or industry experience in the above hypotheses is partly that they serve as proxies for knowledge. For example, both Landy and Vasey (1991) and Tross and Maurer (2000) suggested that one explanation for why longer tenured incumbents gave different job analysis ratings was because they perform tasks differently due to more knowledge. Landy and Vasey (1991) also found some small differences in job analysis ratings by educational level, which is related to knowledge. More directly, job analysis research has generally found that raters unfamiliar with the job being rated do not provide the same job analysis ratings because they lack knowledge of the job (Cornelius, Denisi, & Blencoe, 1984; Harvey & Lozada-Larsen, 1988). Therefore, although previous research has not used SME self-reported knowledge of the job as a predictor of job analysis ratings, related research suggests that more knowledge of the job should lead to better job analysis ratings.

Hypothesis 5: The relationship between importance ratings of personality constructs and observed validities will be stronger when importance ratings are captured from those who report knowing the job extremely well as compared to those reporting knowing it less well.

Data scrubbing and accuracy

As previously discussed, the accuracy of job analysis ratings is usually assessed via measures of agreement or reliability. As with any other measure, both agreement and reliability will be adversely impacted by the presence of outliers. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population (Aguinis *et al.*, 2013; Orr *et al.*, 1991; Roth & Switzer, 2002). What is considered 'abnormal' is somewhat arbitrary and typically left up to the researcher. Biddle (2009) argues for the deletion of any rating that is 1.645 standard deviations above or below the mean, which eliminates any rating above the 95th or below the fifth percentiles (note that he calls for the elimination of individual ratings, not raters). Eliminating potentially aberrant ratings at either extreme should improve the accuracy of the resulting importance ratings.

Hypothesis 6: The relationship between importance ratings of personality constructs and observed validities will be stronger when importance ratings are computed on data from which outliers (ratings 1.645 standard deviations from the mean) have been removed than when importance ratings are computed on the entire sample.

These hypotheses were tested in four samples of validation studies using importance ratings of personality traits collected from job analysis SMEs as independent variables and the validities of those traits in predicting performance as the dependent variables. In each of the samples, validities were accumulated meta-analytically to serve as the criterion or measure of accuracy. Within each sample, SMEs were divided into groups and the importance ratings of each group related to observed validities to test the moderator hypotheses.

Methods

Participants

There were three different types of participants in this study. First were those who participated in the job analysis phase of 48 validation studies, which included 1,533 employees in 25 organizations. Although race, gender, and age were typically not collected from participants during the job analysis, data on their experience (job and industry), role (incumbent or supervisor), frequency of contact with the job, and knowledge of the job were captured. The second and third groups of participants, respectively, were the 12,210 employees and their supervisors who participated in the 48 concurrent validation studies across four job families providing the criterion in this research.

Measures

First, job analysis participants rated the importance of various personality traits for performance in a particular job. Second, these data were aggregated by group to test moderator hypotheses. Third, validities of each trait were obtained from validation studies of these traits by job family. Finally, these meta-analytic validities were used as measures of accuracy to test hypotheses 1–6. Thus, the first set of participants described above, the job analysis SMEs, provided the data used to calculate the independent variables in this study. The second and third sets of participants, from the validation studies, provided the validity data that comprised the dependent variables.

From the job analysis participants, importance ratings were captured on 17 to 21 personality traits using a five-point scale (ranging from 1 = not important at all to 5 = important to a very large extent). A single item was used to measure importance for each trait, and all importance ratings were captured via the Internet. The importance ratings demonstrated reasonable inter-rater reliability, with an ICC2 of 0.89 and a range from 0.63 to 0.98. Although the total number of SMEs varied by study, there was an average of 31.94 SMEs per job (*SD* was 26.72). Descriptive statistics for each trait's importance rating are reported in Table 1.

To test the first hypothesis that importance ratings would be related to validities, the mean importance of rating for each trait was computed across all available job analysis SMEs. To test hypotheses two through six, participants were divided into groups and mean importance ratings computed within group and compared to validities as described below (because traits were the observations in this study, it was not possible to use

Table 1. Average importance, validities, and reliabilities across all jobs

	Importance ratings				Validities				
Trait	*k*	N^1	Mean^2	*SD*	N^3	Mean^4	*SD*	Average rxy	Average rxx
Achievement	44	1,469	4.10	0.36	11,656	42.03	3.09	0.17	0.73
Adaptability	45	1,482	4.32	0.25	11,989	39.56	1.58	0.11	0.77
Analytical	25	810	3.84	0.36	6,821	40.69	1.27	0.13	0.86
Compliance	5	83	4.02	1.05	847	42.94	3.20	0.16	0.75
Concern for others	41	1,359	4.00	0.35	11,028	40.89	1.49	0.10	0.78
Cooperation	45	1,441	4.35	0.26	11,600	43.36	2.21	0.12	0.77
Decisiveness	2	38	4.26	0.30	412	38.97	1.77	0.19	0.72
Dependability	48	1,533	4.51	0.21	12,386	42.50	2.07	0.20	0.78
Detail orientation	47	1,515	4.39	0.29	12,210	39.61	2.24	0.20	0.78
Energy	41	1,356	4.00	0.36	11,190	38.44	2.78	0.19	0.77
Independence	17	632	3.84	0.25	5,530	27.07	2.07	−0.05	0.68
Initiative	44	1,384	4.23	0.25	11,731	41.77	1.73	0.27	0.82
Innovation	14	527	3.68	0.33	4,670	40.38	1.38	0.12	0.89
Integrity	37	1,314	4.32	0.53	10,858	39.43	2.09	0.14	0.71
Leadership	36	1,202	4.30	0.46	9,883	40.17	1.85	0.23	0.84
Optimism	7	151	4.13	0.12	1,516	41.84	1.50	0.19	0.77
Persistence	41	1,404	4.08	0.29	11,288	40.54	1.97	0.16	0.75
Persuasion	16	553	3.51	0.37	3,633	34.77	3.34	0.12	0.75
Self-control	45	1,407	4.44	0.24	11,337	38.15	2.55	0.12	0.79
Social orientation	35	1,216	4.06	0.43	10,207	39.55	2.96	0.09	0.73
Stress tolerance	45	1,433	4.26	0.29	11,580	38.65	1.71	0.13	0.81

Notes. *k* = the number of validation studies in which these traits were used; N^1 = the number of SMEs from which importance ratings were captured; Mean^2 = the average importance rating; *SD* = standard deviation of the importance rating; N^3 = the number of participants in the validation studies; Mean^4 = the average trait score; *SD* = standard deviation of the trait score; rxy = sample-weighted mean validity; rxx = sample-weighted mean internal consistency reliability.

regression to test our moderator hypotheses). Job tenure and industry experience were both self-reported on the following scale: 1 = less than 3 months, 2 = 3–6 months, 3 = 7–11 months, 4 = 1–2 years, 5 = 3–4 years, and 6 = 5 years or more (with an overall mean of 4.25 and *SD* of 1.33). To test hypothesis two, regarding job tenure, there was enough variability to conduct a median split on SME job tenure within each of the 48 job analysis samples in the four job families. As a result, in one organization the median may have involved comparing those rating tenure as a 1 or 2 with those rating it 3 or higher, while in another organization the comparison may have been between those rating it 1–3 with 4 and 5. For industry experience, there was far less variance among participants (with an overall mean of 5.32 and *SD* of 0.84) so the two groups were created by comparing those with '5 years or more' (65.6%) with everyone else (34.4%). To test hypothesis four, participants self-reporting as incumbents (66%) were compared with those reporting as supervisors (34%).

Knowledge of the job was rated on the following scale: (1) extremely well, (2) very well, (3) somewhat well, or (4) not at all. Again, due to the skewed distribution, comparisons were made between SMEs knowing the job extremely well (69.2%) and those reporting anything less (30.8%). Correlations were computed between these variables and are reported in Table 2 (the continuous scales, rather than the dichotomized

Table 2. Descriptive statistics and intercorrelations between independent variables

	Mean	*SD*	2	3	4	5
1. Role	0.70	0.42	0.04	−0.14**	0.06*	−0.01
2. Position tenure	4.20	1.20	1.00	0.38**	0.03	−0.12**
3. Industry experience	5.32	0.84		1.00	0.05*	−0.08**
4. Contact frequency	1.52	0.81			1.00	0.15**
5. Knowledge of job	1.64	0.50				1.00

Notes. Role coded 1 = incumbent, 0 = supervisor; N = 1,533.
*$p < .05$, **$p < .01$.

versions, were used in Table 2). As can be seen, supervisors had significantly more industry experience than incumbents. Surprisingly, self-reported knowledge of the job was slightly and negatively related to both job tenure and industry experience. Finally, to evaluate the impact of outliers, ratings of individual attributes were removed if they were more or less than 1.645 standard deviations from the mean within each of the 48 studies. Thus, the mean number of SMEs rating each trait was reduced slightly (with the grand mean across all trait–study combinations dropping from 43 to 40 SMEs per construct rated).

The validities used as the dependent variables in this study were the validities of a set of personality traits. A personality inventory designed around the O*NET work styles taxonomy (Weekley, Ployhart, & Cooper-Hakim, 2005) was used, whole or in part, in all of the validation studies. The resulting measures were the result of a rigorous three-step process: (1) Six SMEs independently sorted a pool of 524 items into the original 17 scales and items having less than 50% agreement were deleted, resulting in a pool of 370 items; (2) data were collected on the 370 items from 1,169 participants and used to create 12-item scales by retaining those items showing high construct loads and low cross-loads; and (3) data were collected on the remaining 204 items from a new sample of 520 participants and used to create the final 10-item scales again by retaining those items showing high construct loads and low cross-loads. Reliabilities for the majority of these scales were in the mid 1980s, and overall, the items showed reasonable fit with the *a priori* factor structure.

Since originally developed, measures of four additional traits were created and used in a number of the studies (persuasion, compliance, risk-taking, and optimism were added to the original taxonomy of achievement, adaptability, analytical thinking, concern for others, cooperation, dependability, detail orientation, energy, independence, initiative, innovation, integrity, leadership, persistence, self-control, social orientation, and stress tolerance). In each of the validation studies, incumbents completed an experimental test battery containing the personality inventory, whole or in part. Each of the traits was measured by 10 items using a five-point, strongly agree–strongly disagree, scale (e.g., adaptability = 'I prefer things to stay the same and not change'; concern for others = 'I am deeply moved by others' misfortunes'; and innovation = 'I am always coming up with new ideas'). After reverse scoring appropriate items, scale scores were created by summing the 10-item ratings (the online administration system did not allow skipping items, so all 10 items were rated in each case). Internal consistency reliabilities (alpha) for these personality scales across the 48 studies ranged from 0.47 to 0.93 (mean = 0.77 and SD = 0.08).

Independent ratings of performance were captured, almost always on research-only rating forms, from the participants' immediate supervisor. While all 48 studies included a measure of task performance, the number of tasks rated varied considerably across jobs. Additionally, a measure of organizational citizenship behaviour, or OCB, was usually captured using a 10-item scale adapted from (Smith, Organ, & Near, 1983). Finally, where job appropriate, customer service behaviour was also rated on a standard 10-item scale developed by the consulting firm conducting the validation studies. To create an overall performance composite, all available measures (up to three) within a validation study were unit-weighted and averaged. The internal consistency reliabilities for this composite ranged from 0.70 to 0.98 (mean = 0.92 and *SD* = 0.05), suggesting they could be combined into a reliable measure of overall performance. The data included 696 different validity coefficients, an average of 13.98 traits across 48 studies, with validities ranging from −0.19 to 0.59 (mean = 0.11 and *SD* = 0.12). Because the intent is to look at the accuracy of job analysis importance ratings rather than operational validity, the validities were corrected for unreliability in *both* the predictor and criterion. Although not usually done in operational validation studies, corrections for predictor unreliability were made in this case, using the observed coefficient alpha, to eliminate the impact of variance across the predictors in reliability on the job analysis–validity relationship. Because inter-rater reliability data were not available (only one rater was used in every study reported herein), corrections for criterion unreliability were made for illustration using the estimate of 0.52 from the meta-analysis of Viswesvaran, Ones, and Schmidt (1996). Probably because the same correction for criterion unreliability was applied to all coefficients, the results were same as when no correction was applied.

Validities derived from local studies are known to reflect substantial errors in estimation, largely due to sampling error. Meta-analysis, by accumulating results across multiple studies, has proven able to generate more stable and accurate estimates of the strength of relationship between variables. In the current study, there were four job families for which a fair number of validation studies were available: retail associate (k = 12), retail manager (k = 15), customer service representative (k = 10), and sales professional (k = 11). Within job family, sample size-weighted average validities were computed across the corrected validities (as described above) and used as the criteria to test the hypothesis. Conducting analyses within job family enabled four independent tests of the hypotheses using more accurate estimates of validity. Because criterion-related validities were the dependent variable in this study (our operationalization of accuracy), it was important to eliminate controllable sources of variation in validities such as those due to differences in job requirements, which analyses by job family helped achieve. See Table 3 for a depiction of the resulting data sets, one for each of the four job families.

Analyses

To test hypothesis one, the overall mean importance rating for each personality trait was used as the independent variable and the meta-analytic validity for that trait used as the dependent variable *within job family*. Related to hypotheses two through five, mean importance ratings for each trait were computed for (1) low and high job tenure groups, (2) low and high industry experience groups, (3) incumbents versus supervisors, (4) SMEs having daily contact with the job versus anything less, and (5) SMEs reporting knowing the job extremely well versus anything else. To test hypothesis six, the overall mean importance rating used to test hypothesis one was compared to an alternative mean importance rating computed after omitting outliers (as described above). To test

Table 3. Description of the data set used to test all hypotheses

Trait	A Job family validity	B Overall importance rating	C High job tenure importance	D Low job tenure importance	E Incumbent importance	F Supervisor importance	G etc.
1							
2							
3							
i							

Notes. 1. This data set was replicated four times, once for each job family.
2. Column A are the meta-analytically derived estimates of the validity of each personality trait within job family.
3. Column B are the sample-weighted mean importance ratings for each personality trait within job family.
4. Columns C and D are the same data reported in Column B, except broken into high versus low job tenure groups.
5. Columns E and F are the same data reported in Column B, except broken into incumbent versus supervisor groups.
6. In a similar fashion, importance ratings were computed separately for high–low industry experience groups, for groups reporting knowing the job extremely well versus all others, and for the importance ratings after removing outliers (which were compared to column B).
7. Hypothesis 1 was tested by correlating columns A and B.
8. Hypothesis 2 was tested by comparing the correlations between A–C and A–D.
9. Hypothesis 4 was tested by comparing the correlations between A–E with A–F.
10. Hypotheses 3, 5, and 6 were tested with the columns described in note 6 above.

hypotheses two through six, correlations were computed between the average importance ratings, by group, and the scale validities across available work styles measures. This was repeated four times, once within each of the four job families. The number of personality traits on which the correlations were computed varied across the 48 original validation studies from 10 to 21. As a result, the number of traits on which the job analysis–validity correlations were based varied by job family as follows: retail associate (N = 20), retail manager (N = 21), customer service representative (N = 16), and sales professional (N = 18). It is important to emphasize that it is the number of traits which comprise the N or sample size used to test each hypothesis. While much larger numbers of participants contributed the data making up the independent (trait importance ratings) and dependent (trait validities) variables, the 'observations' in the current analysis are the traits and the analyses are replicated four times (by job family). Thus, our samples ranged in size from 16 to 21.

Results

Table 1 shows the descriptive statistics across all of the traits included in the studies reported herein. To test the hypothesis that importance ratings would be related to validities, the correlations between personality importance ratings and criterion-related validities were computed at the job family level (see Table 4a–d). The relationship between importance ratings and validities (see Table 5) was strong in three of the four job families: retail associate (N = 20, r = .49, p < .05), retail manager (N = 21, r = .50, p < .05), and customer service representative (N = 16, r = .56, p < .05). There was no

Table 4. Average importance, validities, and reliabilities for the (a) Customer Service Representative, (b) Retail Associate, (c) Retail Manager, (d) Sales Professional Job Family

	Importance ratings				Validities				
Trait	*k*	*N*[1]	Mean[2]	*SD*	*N*[3]	Mean[4]	*SD*	Average rxy	Average rxx
(a) Customer service representative									
Achievement	8	267	3.75	0.29	1,410	38.25	5.18	0.16	0.76
Adaptability	10	301	4.29	0.32	1,702	37.86	3.11	0.17	0.79
Analytical	3	61	3.64	0.50	425	40.30	1.83	0.17	0.86
Concern for others	10	301	3.81	0.50	1,702	39.67	2.78	0.14	0.77
Cooperation	10	301	4.31	0.27	1,702	41.19	3.88	0.17	0.76
Dependability	10	301	4.42	0.14	1,702	39.66	3.68	0.21	0.77
Detail orientation	10	301	4.40	0.21	1,702	37.41	3.98	0.20	0.75
Energy	6	226	3.54	0.33	1,219	34.24	1.92	0.21	0.78
Initiative	9	222	3.82	0.39	1,582	38.40	3.57	0.26	0.81
Integrity	7	248	4.19	0.47	1,381	37.97	2.19	0.15	0.75
Leadership	5	126	4.37	0.36	926	34.71	4.24	0.23	0.88
Persistence	9	289	3.76	0.32	1,572	38.28	3.78	0.17	0.76
Persuasion	3	163	2.66	0.17	769	31.31	1.17	0.08	0.73
Self-control	10	301	4.47	0.26	1,702	36.67	3.73	0.19	0.78
Social orientation	4	100	4.50	0.27	765	36.23	6.97	0.21	0.76
Stress tolerance	10	301	4.16	0.36	1,702	36.67	3.23	0.16	0.81
(b) Retail associate									
Achievement	10	273	4.17	0.38	3,307	39.85	3.95	0.14	0.70
Adaptability	11	285	4.24	0.33	3,569	38.78	1.58	0.13	0.75
Analytical	5	96	3.90	0.31	1,315	37.25	1.81	0.15	0.85
Compliance	3	49	3.63	1.61	518	42.03	4.57	0.20	0.76
Concern for others	11	285	4.07	0.40	3,569	40.57	1.27	0.13	0.76
Cooperation	12	303	4.55	0.21	3,745	43.21	2.48	0.17	0.73
Dependability	12	303	4.58	0.23	3,745	42.28	2.70	0.18	0.75
Detail orientation	11	285	4.48	0.27	3,569	39.16	1.91	0.18	0.74
Energy	10	224	4.08	0.41	3,159	37.55	4.53	0.24	0.73
Independence	3	102	3.59	0.17	1,341	27.64	2.56	−0.04	0.67
Initiative	11	288	4.06	0.25	3,492	41.25	1.54	0.24	0.79
Innovation	2	49	3.73	0.06	925	39.74	1.11	0.11	0.90
Integrity	11	285	4.24	0.70	3,569	38.83	2.02	0.17	0.69
Leadership	7	191	4.14	0.63	2,285	37.52	2.14	0.17	0.85
Optimism	5	117	4.13	0.15	1,187	40.94	1.75	0.19	0.77
Persistence	8	228	3.98	0.19	3,057	39.61	2.10	0.16	0.73
Persuasion	2	59	3.66	0.32	506	36.85	6.81	0.04	0.78
Self-control	11	250	4.48	0.27	3,329	38.23	2.60	0.17	0.76
Social orientation	10	281	4.04	0.50	3,301	37.37	3.26	0.07	0.70
Stress tolerance	10	232	4.24	0.30	3,153	37.71	1.29	0.15	0.81
(c) Retail manager									
Achievement	15	504	4.26	0.26	4,696	44.37	2.05	0.18	0.74
Adaptability	15	504	4.44	0.16	4,696	41.08	0.80	0.08	0.77
Analytical	12	378	4.01	0.21	3,772	42.15	1.16	0.12	0.87
Compliance	2	34	4.57	0.24	329	44.37	1.05	0.10	0.74
Concern for others	14	489	4.02	0.20	4,626	41.63	1.17	0.06	0.81
Cooperation	14	485	4.26	0.24	4,543	44.38	1.12	0.08	0.80

Continued

Table 4. (*Continued*)

	Importance ratings				Validities				
Trait	*k*	N^1	Mean^2	*SD*	N^3	Mean^4	*SD*	Average rxy	Average rxx
Decisiveness	2	38	4.26	0.30	412	38.97	1.77	0.19	0.72
Dependability	15	504	4.52	0.22	4,696	43.59	1.15	0.22	0.80
Detail orientation	15	504	4.33	0.33	4,696	40.79	1.90	0.24	0.81
Energy	15	504	4.27	0.29	4,696	40.23	1.79	0.17	0.79
Independence	9	291	3.62	0.32	3,199	25.79	2.24	−0.04	0.70
Initiative	15	504	4.49	0.13	4,696	43.37	1.24	0.29	0.84
Innovation	10	305	3.66	0.34	3,240	40.67	1.56	0.13	0.89
Integrity	12	456	4.44	0.46	4,432	41.31	2.17	0.11	0.71
Leadership	15	504	4.57	0.30	4,696	43.00	1.16	0.25	0.80
Optimism	2	34	4.14	0.01	329	45.11	0.58	0.19	0.78
Persistence	14	485	4.26	0.24	4,543	42.03	1.49	0.17	0.75
Persuasion	4	114	4.05	0.31	866	37.87	4.08	0.14	0.74
Self-control	15	504	4.46	0.15	4,696	39.12	1.89	0.08	0.82
Social orientation	13	466	4.08	0.39	4,284	41.13	2.81	0.09	0.75
Stress tolerance	15	504	4.35	0.21	4,696	40.11	1.45	0.11	0.82
(d) Sales professional									
Achievement	11	425	4.09	0.50	2,243	42.74	2.66	0.22	0.75
Adaptability	9	392	4.23	0.24	2,022	38.86	2.10	0.10	0.76
Analytical	5	275	3.62	0.55	1,309	40.04	0.85	0.13	0.83
Concern for others	6	284	4.08	0.40	1,131	40.68	1.60	0.15	0.75
Cooperation	9	352	4.32	0.31	1,610	43.14	2.86	0.03	0.76
Dependability	11	425	4.52	0.25	2,243	42.76	1.74	0.17	0.77
Detail orientation	11	425	4.40	0.32	2,243	39.53	2.13	0.17	0.78
Energy	10	402	3.86	0.43	2,116	38.22	2.85	0.17	0.80
Independence	4	214	4.26	0.22	831	29.18	1.02	−0.08	0.68
Initiative	9	370	4.26	0.32	1,961	41.59	1.76	0.26	0.83
Innovation	2	173	3.70	0.38	505	39.70	0.69	0.03	0.90
Integrity	7	325	4.31	0.52	1,476	36.59	1.94	0.12	0.71
Leadership	9	381	4.00	0.61	1,976	39.06	2.05	0.25	0.88
Persistence	10	402	4.14	0.38	2,116	40.36	1.48	0.15	0.76
Persuasion	7	217	3.83	0.56	1,492	34.04	2.84	0.16	0.76
Self-control	9	352	4.35	0.33	1,610	36.75	3.10	0.08	0.76
Social orientation	8	369	3.93	0.48	1,857	41.15	1.13	0.07	0.74
Stress tolerance	10	396	4.22	0.35	2,029	38.42	1.67	0.10	0.81

Note. *k* = the number of validation studies in which these traits were used; N^1 = the number of SMEs from which importance ratings were captured; Mean^2 = the average importance rating; *SD* = of the importance rating; N^3 = the number of participants in the validation studies; Mean^4 = the average trait score; *SD* = of the trait score; rxy = sample-weighted mean validity; rxx = sample-weighted mean internal consistency reliability.

relationship between the two for the sales professional job family ($N = 18, r = -.02, ns$). The simple mean relationship between the job analysis importance ratings and validities across the four job families was $r = .38$. Because some of the validity coefficients used as criteria in these studies were based on relatively small samples, making them potentially less useful measures of accuracy, the analyses described above were rerun after omitting traits with an *N* for the validity estimate below 600. The relationships strengthened

Table 5. Importance–validity correlations by job family

Job family	*k*	*N*	*r*	*p* one-tailed
Retail associate	12	20	.50	<.05
Retail manager	15	21	.49	<.05
Customer service representative	10	16	.56	<.05
Sales professional	11	18	−.02	ns

slightly as follows: retail associate ($N = 18$, $r = .55$, $p < .05$), retail manager ($N = 18$, $r = .59$, $p < .05$), customer service representative ($N = 15$, $r = .56$, $p < .05$), and sales professional ($N = 16$, $r = -.10$, *ns*), and the simple mean relationship between the job analysis importance ratings and validities across the four job families increased to $r = .40$. Thus, in three of the four studies there was a strong and significant relationship between the importance ratings provided by SMEs of various traits and the subsequent validities observed when measures of those traits were compared to measures of performance. Thus, hypothesis 1 was generally supported.

Hypothesis two that job tenure would moderate the relationship between importance ratings of constructs and observed validities was not supported. Table 6 shows the importance–validity relationships by tenure group. In one case, importance ratings from higher tenure SMEs were more strongly related to validity than those from less experienced SMEs, but the exact opposite was found in two other cases and the average across the four job families was not significantly different ($r = .33$ for high-tenure SMEs and .40 for less tenured SMEs). Hypothesis three that industry experience would moderate the job analysis–validity relationship was also not supported. As shown in Table 7, only one of the comparisons was statistically significant and that difference actually favoured those with less industry experience. Thus, experience, be it with the job or within the industry, does not appear to impact the accuracy with which SMEs report the importance of personality traits.

Table 6. Comparison of high and low job-tenured job analysis raters

Job family	High tenure	Low tenure	*Z*
Sales	0.03	0.32	2.99, $p < .05$ (one-tailed)
Customer service representative	0.22	0.45	2.22, $p < .05$ (one-tailed)
Retail associate	0.54	0.48	0.67, ns
Retail manager	0.54	0.35	2.35, $p < .01$ (one-tailed)
Average	0.33	0.40	0.64, ns

Table 7. Comparison of high and low industry experienced job analysis raters

Job family	High industry	Low industry	*Z*
Sales	0.04	0.19	1.20, ns
Customer service representative	0.26	0.62	2.69, $p < .05$ (one-tailed)
Retail associate	0.46	0.45	0.12, ns
Retail manager	0.53	0.62	1.05, ns
Average	0.32	0.47	1.49, ns

Table 8. Comparison of incumbent and supervisor job analysis raters

Job family	Incumbents	Supervisors	*Z*
Sales	0.16	0.28	0.89, ns
Customer service representative	0.15	0.48	2.66, $p < .01$ (one-tailed)
Retail associate	0.36	0.61	2.53, $p < .01$ (one-tailed)
Retail manager	0.50	0.35	1.67, $p < .05$ (one-tailed)
Average	0.29	0.43	1.68, $p < .05$ (one-tailed)

Hypothesis four that importance ratings will be more strongly related to validities when captured from supervisors than from incumbents was partially supported. Table 8 shows the importance–validity relationships by job role. In two job families, importance ratings from supervisors were significantly more strongly related to validity than those from incumbent SMEs, whereas the relationship was reversed in one case and not significant in the other. Although not large, the *average* difference ($r = .29$ for incumbents vs. .43 for supervisors) was significant, with supervisors providing slightly more accurate ratings of the job relatedness of personality constructs than incumbents.

Hypothesis five that SMEs reporting to know the job 'extremely well' would provide more accurate importance ratings than those reporting anything less was also partially supported. In all four samples (see Table 9), the results were in the predicted direction. The differences were statistically significant in two of the samples and for the *average* values computed across the four samples ($r = .40$ for those knowing the job extremely well vs. $r = .22$ for anything less). Thus, deleting the relatively small proportion of SMEs reporting less than the highest level of knowledge of the job did increase the accuracy of the job analysis ratings captured.

Finally, hypothesis six that deleting outliers would improve accuracy was not supported. As indicated in Table 10, removing outliers had no impact on the job analysis–validity relationship (e.g., the averages were 0.34 and 0.33). While omitting extreme

Table 9. Comparison of high and low knowledge of job analysis raters

Job family	Extremely well	Everything else	*Z*
Sales	0.13	−0.02	1.20, ns
Customer service representative	0.34	0.06	1.72, $p < .05$ (one-tailed)
Retail associate	0.53	0.36	2.95, $p < .01$
Retail manager	0.58	0.48	1.33, $p < .01$ (one-tailed)
Average	0.40	0.22	1.67, p .05 (one-tailed)

Table 10. Comparison with outliers removed

Job family	Removed outliers	Total sample	*Z*
Sales	0.02	−0.02	0.52, ns
Customer service representative	0.29	0.26	0.25, ns
Retail associate	0.55	0.53	0.38, ns
Retail manager	0.49	0.57	1.33, $p < .10$ (one-tailed)
Average	0.34	0.33	0.09, ns

ratings makes some intuitive sense, these data suggest that basing job analysis judgements on the entire sample of SMEs is just as accurate.

In summary, looking at the average relationships observed across the four job families we conclude that (1) importance ratings are related to validities, (2) supervisors provide more accurate importance ratings than incumbents, and (3) those who report knowing the job extremely well also provide more accurate importance ratings than those reporting anything less. Unsupported were the hypotheses that job tenure, industry experience, or outlier deletion would moderate the importance ratings–validity relationship.

Discussion

The primary contribution of this research is the illustration of the use of test validity as a measure of accuracy, providing a new lens with which to evaluate job analysis practices. A secondary contribution is the affirmation of the relevance of content validity to criterion-related validity. Be they based on importance ratings, time spent ratings, or something else, assessment developers have for decades made the leap of faith from SME judgement to test construction and validation. The current research adds to the very slim body of evidence that this leap of faith has been justified. Across the four samples, the average relationship between rated importance and validity was in the 0.4 range.

Evaluating the magnitude of this relationship is difficult because there is so little with which to compare it. In an unpublished study from a single sample, Weekley *et al.* (2013) found a correlation of 0.41 between SME ratings of skill relevance and the validity of skill training experience in predicting performance. Two other studies reported the data necessary to replicate the analyses testing hypothesis one. Robertson and Kinder (1993) reported 33 validities and importance ratings for a variety of personality traits, from which we computed a correlation between them of $r = .38$ ($p < .05$). Similarly, Highhouse, Zickar, Brooks, Reeve, and Sarkar-Barney (2016) reported job analysis ratings and validities for 12 traits. The correlation we computed between these by criterion was as follows: −0.31 (absenteeism), 0.42 (service), and −0.53 (discipline). Ignoring the sign for the two negatively keyed criteria, the average across the three in the Highhouse *et al.* (2016) research was 0.42. In short, the job analysis–validity results generated by analysing data reported by both Robertson and Kinder (1993) and Highhouse *et al.* (2016), plus the previous Weekley *et al.* (2013) study, are nearly identical to what was found in the current studies – a correlation between construct importance ratings and validities in the 0.4 range. Cohen's (1988) rules of thumb would suggest that this is a moderately strong relationship. More recent analyses, however, have provided evidence that much smaller relationships (e.g., $r = .30$) may in fact be 'large' (Bosco, Aguinis, Singh, Field, & Pierce, 2015).

Implications for theory

Tested was a new way to operationalize the concept of accuracy in job analysis ratings by SMEs – the criterion-related validities of tests. Consistent with the suggestions of Morgeson and Campion (2000) and Sanchez and Levine (2000), the accuracy of the inferences drawn was established through comparison with the outcomes the job analysis data were collected to support. As opposed to measures of agreement or reliability, using actual outcomes such as validity should provide greater clarity as to which job analysis

practices result in the most accurate data and most effective processes. The job analytic practices producing the highest construct importance–validity relationships would be the most accurate. The often noted absence of an overarching theory in job analysis (e.g., Peterson *et al.*, 2001a; Sanchez & Levine, 2012) may be due in part to an inability to test it. By filling this void, the current research will hopefully stimulate greater theory building in the realm of job analysis. The use of outcomes as criteria or measures of accuracy is not limited to the selection context. Using performance difference scores ('d') between groups, for example, would be a straightforward way of comparing the efficacy of different training methods and even approaches to training development that is based on job analysis.

Showing a direct relationship between SME judgements, which are commonly used to estimate content validity, and actual criterion-related validity also supports the unified view of validity. Under the unified concept of validity, what were once described as different types of validity (i.e., construct, content, and criterion-related validity) came to be considered different forms of evidence of validity (Landy, 1986). According to this point of view, the accuracy of the inference drawn from a predictor depends on its relationship the criterion measures, how accurately the predictor and criterion measures capture their intended domains, and how accurately the psychological domain measured by the predictor was deduced from the performance domain (Binning & Barrett, 1989). This latter linkage has typically been established via job analysis, wherein SMEs and job analysts are assumed to be able to identify the individual attributes differentiating high performers from the rest. Content validity requires the establishment of links between the content of the job and the content of the test. Be they based on importance ratings, relevance ratings, time spent ratings, or something else, industrial/organizational psychologists have for decades made the leap of faith from SME judgement to test construction and validation.

Murphy (2009) stimulated a spirited debate (see also the 12 commentaries to his focal article in that same issue) and directly challenged the unified view of validity, with his assertion that content validation is often unrelated to criterion-related validity. Of the three methods for demonstrating content validity described by Murphy (2009), the linkage method used in the current study is probably the most common. In this method, linkages are made judgementally between the KSAO constructs and both the job performance domain and test content. The extent of overlap determines the appropriateness of the test for use in selection. As with all job analysis methods, this form of content validation is predicated on judgements of unknown accuracy. If the unified concept of validity is correct, then content accurately judged to be more important or relevant to the job performance domain should show higher criterion-related validities than those judged less important or irrelevant. The current research adds to the very slim body of evidence that this leap of faith has been justified and affirms the relevance of content validity to criterion-related validity and the unified view of validity.

Implications for practice

The results indicate that supervisors and SMEs who know the job extremely well may provide more accurate job requirements information than incumbents or those knowing the job less well, respectively (although the use of either should result in a valid assessment battery). Thus, practitioners would be advised to include in their sample of SMEs a healthy dose of supervisors of the job being analysed. This appears to be true even if the supervisor did not previously hold the job, questioning the long-held belief that 'expertise' comes only from having performed the job. Although requiring replication,

asking SMEs how well they know the job and omitting those who are not supremely confident in their knowledge is an easy and low-cost means of potentially improving the accuracy of the data captured.

Conversely, neither job tenure nor industry experience were related to the accuracy of the job analysis information collected, at least not at the levels studied herein. However, all SMEs were selected because they were deemed to have been on the job long enough to know it reasonably well. We suspect this is typical of job analyses in practice and indicates that, beyond the minimal level required by each job's learning curve, additional experience on the job or in the industry is unnecessary to report accurately on its requirements, that is, the levels of experience with the job needed to report accurately on its requirements may be quite low. Finally, there was nothing in the results to suggest eliminating outliers (observations above the 95th percentile or below the fifth percentile) from the job analysis importance ratings is advisable.

Finally, this study demonstrated that job analysis ratings can be related to the outcomes they were designed to support, in this case test validities. These results are important because they confirm the efficacy of current practice. Those conducting criterion-related validation research will be more likely to observe expected relationships when those expectations are based on the results of job analysis. More importantly, where criterion-related validity cannot be established (e.g., because of small samples), practitioners can have greater confidence in the validity of assessments deployed on the basis of a job analysis alone. The one caveat to this, of course, is the unknown effect of autonomy or job crafting on the job analysis–validity relationships. In highly autonomous jobs, where incumbents are allowed great latitude in how they go about accomplishing objectives, job analysis mean importance ratings may provide far less insight into the validity of the constructs so rated. In jobs where there is more than one way to be successful (e.g., a sales job in which goals can be met either through growing current accounts – 'farming' – or through finding entirely new accounts – 'hunting'), mean importance ratings may obscure important differences.

Limitations and future research

There are several limitations that must be recognized. All of the various personality traits received relatively high importance ratings (the mean across all traits and studies was 4.22 on a five-point scale), suggesting that these SMEs had a difficult time making job-related discriminations on personality traits. The limited variance in the importance ratings undoubtedly reduced the observed correlations. Variance in importance ratings of the personality traits was further reduced by excluding low-rated traits in some of the criterion validation studies. Although the entire inventory was used in many cases, the traits with the lowest mean importance ratings were not included in the subsequent validation study in some instances. Restriction of range in the independent variable means that the magnitude of the relationships observed herein may be conservative.

Even after accumulating results to the job family level, the importance–validity relationship did not always hold up (i.e., in the sales job family). A review of the 11 constituent jobs revealed a wide variety in content (selling products vs. services), price (transactions of $100 to $100,000 per sale), and approach (consultative multiple-contact sales vs. 'one-contact' transactional sales). Further, some were 'farmers' (selling into existing clients), some 'hunters' (selling to new prospects), and others a blend of both. This may have been too broad of a set of jobs to be considered a job family. A third limitation was the exclusive reliance on personality measures. The job relevance of constructs such as knowledge and skill may be more easily rated, which in turn could

result in higher observed importance–validity relationships for those construct domains. While some indirect efforts to link job analysis ratings to cognitive ability validities have been unsuccessful, most have been in military samples using a highly intercorrelated test battery (the ASVAB) and jobs that probably do not vary widely in cognitive requirements (Earles & Ree, 1992; Fairbanks *et al.*, 1990; Jones & Ree, 1998; Peterson *et al.*, 2001a).

There are numerous avenues for future research. As suggested above, extensions of this line of inquiry to other constructs, such as mental abilities, and use of other types of measures, such as objectively scored multiple choice tests, would be helpful in confirming that construct importance ratings are related to validity. As there is some evidence that importance ratings of knowledge and skills are more consistent across SMEs than are measures of personality (Borman, Kubisiak, & Schneider, 1999; Tsacoumis & Van Iddekinge, 2006), there exists the potential that the validity of these is better predicted from a job analysis than constructs such as personality. Likewise, the dependent variable in this paradigm, criterion-related validities, was all based on self-reported measures of personality. Objectively scored measures of knowledge, skills, or abilities may yield different results and should be explored.

Just as different constructs should be examined, the accuracy of different types of job analysis scales can be compared. Scales that assess 'differentiation' or 'discriminability' (the extent to which the construct differentiates high performers from others) may be more useful than importance ratings. This is consistent with the emphasis in competency modelling on the facets that differentiate high performers from the rest (Campion *et al.*, 2011). In one example in the authors' experience, numerical reasoning was rated as important for professional accountant jobs, but SMEs rated it low on its ability to differentiate high from low performers. Given that this ability is uniformly high in most hires from top university accounting programmes, the SMEs recognized the resulting range restriction. Similarly, perhaps relative importance ratings or rankings would improve the discriminability of the data. This is especially the case with personality traits because they all sound so desirable (Raymark, Schmit, & Guion, 1997). Thus, approaches to expand variance in the importance ratings of the personality traits (e.g., Aguinis, Mazurkiewicz, & Heggestad, 2008; Raymark *et al.*, 1997) may result in stronger relationships with the criterion of test validity.

Just as scales should be compared, so too should SMEs. While there is some evidence that experienced SMEs focus on different features of jobs than do less experienced ones (e.g., Borman *et al.*, 1992; Landy & Vasey, 1991), we found no evidence in the current research that more experienced SMEs were more accurate. While common sense would suggest the more experienced SMEs to be more accurate, it may be the case that less experienced ones report more accurately on those attributes actually needed at hire. As it has been recommended that selection procedures focus only on those constructs needed at hire, research comparing wider ranges of experience on the needed-at-hire construct may find important differences in accuracy.

More research comparing the relative accuracy of information collected from incumbents and supervisors is warranted. While far from conclusive, the results suggest that job requirement information captured from supervisors may be more accurate than that coming from incumbents. In addition to confirming this, research into the reasons why would be a valuable contribution. While Morgeson *et al.* (2004) found evidence of self-presentation bias among incumbents but not supervisors, further consideration of this potential bias is needed. Finally, would the same results be observed in jobs offering a higher level of autonomy to incumbents (and greater potential for job crafting)? Most of

the jobs studied here could be described as highly scripted (e.g., call centre representative), allowing little opportunity for job crafting. And while managers and sales representatives would presumably have higher levels of autonomy in how they go about doing their jobs, all of the jobs studied here were at the lower levels of those professions. It is unknown how much latitude they actually had in determining where to focus or how to go about performing their jobs. It remains a possibility that job crafting enables multiple paths to success, with multiple requirement profiles. If true, an average importance rating taken across SMEs would seemingly blur these distinctions and result in less accurate specifications of key requirements.

Acknowledgements

The authors would like to thank Juan Sanchez for his many helpful comments on this paper.

References

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2008). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*, 405–438.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*, 270–301.

Ang, S., Slaughter, S., & Ng, K. Y. (2002). Human capital and institutional determinants of information technology compensation: Modeling multilevel and cross-level interactions. *Management Science*, *48*, 1427–1445. https://doi.org/10.1287/mnsc.48.11.1427.264

Befort, N., & Hattrup, K. (2003). Valuing task and contextual performance: Experience, job roles, and ratings of the importance of job behaviors. *Applied Human Resources Management*, *8*, 17–32.

Biddle, D. A. (2009). *Guidelines oriented job analysis (GOJA): A job analysis process for selection procedure development and validation*. Folson, CA: Biddle Consulting Group.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494. https://doi.org/10.1037/0021-9010.74.3.478

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

Borman, W. C., Kubisiak, U. C., & Schneider, R. J. (1999). Work styles. 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 the O*NET* (pp. 213–226). Washington, DC: American Psychological Association.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. https://doi.org/10.1037/a0038047

Bruning, P., & Campion, M. A. (2018). A role-resource approach-avoidance model of job crafting: A multi-method integration and extension of job crafting theory. *Academy of Management Journal*, *61*, 499–522. https://doi.org/10.5465/amj.2015.0604

Campion, M. A., Fink, A. A., Ruggerberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology*, *64*, 225–262. https://doi.org/10.1111/j.1744-6570.2010.01207.x

Carrier, M. R., Dalessio, A. T., & Brown, S. H. (1990). Correspondence between estimates of content and criterion-related validity values. *Personnel Psychology*, *43*, 85–100. https://doi.org/10.1111/j.1744-6570.1990.tb02007.x

Cascio, W. F., & Aguinis, H. (2008). Research in industrial and organizational psychology from 1963 to 2007. *Journal of Applied Psychology*, *93*, 1062–1081. https://doi.org/10.1037/0021-9010.93.5.1062

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale NJ: Erlbaum.

Cornelius, E. T., Denisi, A. S., & Blencoe, A. G. (1984). Expert and naive raters using the PAQ: Does it matter? *Personnel Psychology*, *37*, 453–464. https://doi.org/10.1111/j.1744-6570.1984.tb00522.x

Cornelius, E. T., & Lyness, K. S. (1980). A comparison of wholistic and decomposed judgment strategies in job analysis by incumbents. *Journal of Applied Psychology*, *65*, 155–163. https://doi.org/10.1037/0021-9010.65.2.155

Dierdorf, E. C., & Morgeson, F. P. (2007). Consensus in work role requirements: The influence of discrete occupational context on role expectations. *Journal of Applied Psychology*, *92*, 1228–1241. https://doi.org/10.1037/0021-9010.92.5.1228

Dierdorff, E. C., & Morgesen, F. P. (2009). Effects of descriptor specificity and observability on incumbent work analysis ratings. *Personnel Psychology*, *62*, 601–628.

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology*, *88*, 635–646. https://doi.org/10.1037/0021-9010.88.4.635

Dokko, G., Wilk, S. L., & Rothbard, N. P. (2009). Unpacking prior experience: How career history affects job performance. *Organization Science*, *20*(1), 51–68. https://doi.org/10.1287/orsc.1080.0357

Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment*, *1*, 153–157. https://doi.org/10.1111/j.1468-2389.1993.tb00103.x

Earles, J. A., & Ree, M. J. (1992). The predictive validity of the ASVAB for training grades. *Educational and Psychological Measurement*, *52*, 721–725. https://doi.org/10.1177/0013164492052003022

Fairbanks, B. A., Welsh, J. R., & Sawin, L. L. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Validity of ASVAB Form 14 for predicting high school course grades (Interim report for period May 1988 to May 1990 AFHRL-TR-90-48)*. Brooks Air Force Base. TX: Air Force Human Resources Laboratory.

Goldsmith, A. H., & Veum, J. R. (2002). Wages and the composition of experience. *Southern Economics Journal*, *69*, 429–443. https://doi.org/10.2307/1061681

Goldstein, H., Noonan, K., & Schneider, B. (1992). *Individual, contextual, and performance correlates of job analysis survey responses*. Paper presented at the Seventh Annual Conference of the Society for Industrial and Organizational Psychology, Montreal Canada.

Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology*, *39*, 543–564. https://doi.org/10.1111/j.1744-6570.1986.tb00952.x

Guder, E. J. (2012). Identifying appropriate sources of work Information. In M. A. Wilson, W. Bennenett.. , S. G. Gibson & G. M. Alliger (Eds.), *The handbook of work analysis: Methods, systems, applications, and science of work measurement in organizations*. New York, NY: Routledge.

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

Harvey, R. J., & Lozada-Larsen, S. R. (1988). Influence of amount of job description information on job analysis accuracy. *Journal of Applied Psychology*, *73*, 457–461. https://doi.org/10.1037/0021-9010.73.3.457

Highhouse, S., Zickar, M. J., Brooks, M. E., Reeve, C. L., & Sarkar-Barney, S. T. (2016). A public-domain personality items bank for use with Raymark, Schmit, and Guion (1997). *Personnel and Assessment Decisions*, *2*, 48–56.

Jones, G. E., & Ree, M. J. (1998). Aptitude test score validity: No moderating effect due to job ability requirement differences. *Educational and Psychological Measurement*, *58*, 284–294. https://doi.org/10.1177/0013164498058002011

Jones, R. G., Sanchez, J. I., Parameswaran, G., Phelps, J., Shoptaugh, C., Williams, M., & White, S. (2001). Selection or training? A two-fold test of the validity of job-analytic ratings of trainability. *Journal of Business Psychology*, *15*, 363–389. https://doi.org/10.1023/A:1007804815480

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *America Psychologist*, *41*, 1183–1192.

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. *Personnel Psychology*, *44*, 27–50. https://doi.org/10.1111/j.1744-6570.1991.tb00689.x

Levine, E. L., Ash, R. A., & Bennett, N. (1980). Exploratory comparative study of four job analysis methods. *Journal of Applied Psychology*, *65*, 524–535. https://doi.org/10.1037/0021-9010.65.5.524

Manson, T. M. (2004). *Cursory versus comprehensive job analysis for personnel selection: A consequential validity analysis*. Unpublished doctoral dissertation, University of South Florida, Tampa.

Manson, T. M., Levine, E. L., & Brannick, M. T. (2000). The validity of task inventory ratings: A multitrait-multimethod analysis. *Human Performance*, *13*, 1–22. https://doi.org/10.1207/S15327043HUP1301_1

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730–740. https://doi.org/10.1037/0021-9010.86.4.730

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of performance. *Journal of Applied Psychology*, *73*, 327–330. https://doi.org/10.1037/0021-9010.73.2.327

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, *82*, 627–655. https://doi.org/10.1037/0021-9010.82.5.627

Morgeson, F. P., & Campion, M. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior*, *21*, 819–827. https://doi.org/10.1002/(ISSN)1099-1379

Morgeson, F. P., Delaney-Klinger, K., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks and competencies. *Journal of Applied Psychology*, *89*, 674–686. https://doi.org/10.1037/0021-9010.89.4.674

Morgeson, F. P., & Dierdorf, E. C. (2011). Work analysis: From technique to theory. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology* (pp. 3–41). Washington, DC: American Psychological Association.

Mullins, W. C., & Kimbrough, W. W. (1988). Group composition as a determinant of job analysis outcomes. *Journal of Applied Psychology*, *73*, 657–664. https://doi.org/10.1037/0021-9010.73.4.657

Murphy, K. R. (2009). Content validity is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 453–464.

Murphy, K. R., Dzieweczynski, J. L., & Zhang, Y. (2009). Positive manifold limits the relevance of content-matching strategies for validating selection batteries. *Journal of Applied Psychology*, *94*, 1018–1031. https://doi.org/10.1037/a0014075

Orr, J. M., Sackett, P. R., & Debois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey or researcher beliefs and an empirical illustration. *Personnel Psychology*, *44*, 473–486.

Ostroff, C., & Schmitt, N. (1987). *The relationship between content and criterion-related validity indices*. Paper presented at the 2nd Annual Meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA

Parent, D. (2000). Industry-specific capital and the wage profile: Evidence from the National Longitudinal Survey of Youth and the Panel Study of Income Dynamics. *Journal of Labor Economics*, *18*, 306–323. https://doi.org/10.1086/209960

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., . . . Dye, D. (2001a). Understanding work using the occupational information network (O*NET): Implications for practice and research. *Personnel Psychology*, *54*, 451–492. https://doi.org/10.1111/j.1744-6570.2001.tb00100.x

Peterson, N. G., Wise, L. L., Arabian, J., & Hoffman, R. G. (2001b). Synthetic validation and validity generalization when empirical validation is not possible. In J. Campbell & D. Knapp (Eds.), *Exploring the upper limits of personnel selection and classification*. Mahwah, NJ: Erlbaum.

Prien, K. O., Prien, E. P., & Wooten, W. (2003). Interrater reliability in job analysis: Differences in strategy and perspective. *Public Personnel Management*, *32*, 125–141. https://doi.org/10.1177/009102600303200107

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employees selection. *Personnel Psychology*, *50*, 723–736. https://doi.org/10.1111/j.1744-6570.1997.tb00712.x

Richman, W. L., & Quinones, M. A. (1996). Task frequency rating accuracy: The effect of task engagement and experience. *Journal of Applied Psychology*, *81*, 512–524. https://doi.org/10.1037/0021-9010.81.5.512

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology*, *66*, 225–244. https://doi.org/10.1111/j.2044-8325.1993.tb00534.x

Roth, P. L., & Switzer, F. S. (2002). Outliers and influential cases: Handling those discordant contaminated maverick rogues. In S. G. Rogelberg (Ed.), *Blackwell handbook of research methods in industrial and organizational psychology* (pp. 297–309). Oxford: Blackwell.

Sanchez, J. I. (1990). *The effects of job experience on judgments of task importance*. Paper presented at the Annual Conference for the Society of Industrial and Organizational Psychology, Miami.

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. https://doi.org/10.1002/(ISSN)1099-1379

Sanchez, J. L., & Levine, E. L. (2012). The rise and fall of job analysis and the future of work analysis. In S. T. Fiske (Ed.), *Annual review of psychology* (63, pp. 397–425). Palo Alto, CA: Annual Reviews.

Sanchez, J. I., Prager, I., Wilson, A., & Viswesvaran, C. (1998). Understanding within-job title variance in job-analytic ratings. *Journal of Business Psychology*, *12*, 407–420. https://doi.org/10.1023/A:1025046921171

Schmitt, N., & Cohen, S. A. (1989). Internal analyses of task ratings by job incumbents. *Journal of Applied Psychology*, *74*, 96–104. https://doi.org/10.1037/0021-9010.74.1.96

Smith, A., Organ, D. W., & Near, J. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, *68*, 653–663. https://doi.org/10.1037/0021-9010.68.4.653

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*, 703–742.

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. https://doi.org/10.1023/A:1007770919305

Tsacoumis, S., & Van Iddekinge, C. (2006). *A comparison of incumbent and analyst ratings of O*NET skills*. http://www.onetcenter.org/reports/SkillsComp.html.

Van Iddekinge, C. H., Raymark, P. H., & Eidson, C. E. (2011). An examination of the validity and incremental value of needed-at-entry ratings for a customer service job. *Applied Psychology: An International Review*, *60*, 24–45. https://doi.org/10.1111/j.1464-0597.2010.00425.x

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Weekley, J. A., Hauck, E. L., Burner, T., & Evans, S. (2013). *Accuracy of job analysis ratings in predicting training experience validities*. Paper presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, TX.

Weekley, J. A., Ployhart, R. E., & Cooper-Hakim, A. (2005). *On the Development of Measures of the O*NET Work Styles*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Weisner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and the degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, *61*, 275–290. https://doi.org/10.1111/j.2044-8325.1988.tb00467.x

Wrzesniewski, A., & Dutton, J. E. (2001). Crafting a job: Revisioning employees as active crafters of their work. *Academy of Management Review*, *26*, 179–201. https://doi.org/10.5465/amr.2001.4378011

Received 16 July 2018; revised version received 25 March 2019