Personnel Assessment and Decisions

Volume 5 | Issue 1 Article 6

2019

# Are Consensus Ratings of Functional Job Analysis Scales More Reliable than Ratings Made by Independent Raters?

Greg A. Chung-Yan
*University of Windsor*, gcy@uwindsor.ca

Aaron C. H. Schat
*McMaster University*, schata@mcmaster.ca

Steven F. Cronshaw
*University of Northern British Columbia (Emeritus)*, steven.cronshaw@unbc.ca

## Recommended Citation

Chung-Yan, Greg A.; Schat, Aaron C. H.; and Cronshaw, Steven F. (2019) "Are Consensus Ratings of Functional Job Analysis Scales More Reliable than Ratings Made by Independent Raters?," *Personnel Assessment and Decisions*: Vol. 5 : Iss. 1 , Article 6.
DOI: 10.25035/pad.2019.01.006
Available at: https://scholarworks.bgsu.edu/pad/vol5/iss1/6

# Are Consensus Ratings of Functional Job Analysis Scales More Reliable Than Ratings Made by Independent Raters?

Greg A. Chung-Yan[1], Aaron C. H. Schat[2], and Steven F. Cronshaw[3]

1. Department of Psychology, University of Windsor
2. DeGroote School of Business, McMaster University
3. School of Business, University of Northern British Columbia *(Emeritus)*

KEYWORDS

job analysis, rating reliability, functional Job analysis, consensus ratings, generalizability theory, measurement

ABSTRACT

This study addresses an open research question in regard to a well-established and widely used job analysis system, Functional Job Analysis (FJA): Are consensus ratings of the FJA scales more reliable than the independent scale ratings that are the norm in job analysis application and the related research literature? In our experimental study, we found that this is not the case: No significant difference is found between consensus and independent ratings of the FJA scales. The reasons for this finding are explored as well as its relevance to the validity of the FJA system.  Implications for other work and job analysis systems are discussed.

Many human resources professionals who make personnel assessments and decisions also make the scale ratings that are central to many job analysis systems. The reliability of these scale ratings is important as evidenced by the attention paid by researchers and practitioners to meta-analytic results of the reliability of job analysis ratings (e.g., Dierdorff & Wilson, 2003; Voskuijl & van Sliedregt, 2002). Furthermore, many practitioners and scholars operate under the assumption that job analysis is necessary to support a wide range of personnel interventions, including human resource selection, performance appraisal and management, training, and job evaluation. Nevertheless, despite decades of research on the reliability of job analysis ratings, a question remains unanswered that is addressed in this study: Does the reliability of Functional Job Analysis (FJA) ratings obtained by consensus across multiple raters exceed those made by independent raters?

Typically, investigatory focus is on the reliability (usually interrater) of job analysis ratings made by single raters to the exclusion of consensus ratings.  FJA scale ratings are no exception. The only research on the reliability of FJA scale ratings to date was conducted using single independent raters (Schmitt & Fine, 1983). Especially given that Fine and Getkate (1995) strongly recommend attaining consensus between FJA raters in order to ensure the consistency of FJA scale ratings, it is important to know if consensus ratings using the FJA scales are in fact superior in reliability to ratings by single raters. This study addresses that question.

FJA scales are somewhat novel in job and work analysis. Job analysis reliability research has focused largely on measurements of task frequency, importance, difficulty, and time spent, which are rating scales meant to assess *extrapolated correlates* of job tasks and not the *substantive content* of the tasks themselves. The extrapolated correlates approach requires that the rater extrapolate from task attributes to a metric that is correlated with—but assesses characteristics extrinsic to—the content of the work being performed; these ratings are made primarily for administrative purposes (e.g., time-spent ratings used in work scheduling

Corresponding author:
Greg. A Chung-Yan
Department of Psychology, University of Windsor
Windsor, Ontario, Canada, N9B3P4
Email: gcy@uwindsor.ca

We thank the six FJA raters for their participation in this research.

and time management). The substantive content approach requires the rater to only make judgments about and ratings of the task content (e.g., the complexity of a task in relation to Things, Data, and People).[1] These FJA ratings of substantive content are used primarily for quality control in the collection of the FJA data but can also be used when developing personnel interventions (e.g., selection, job design). It is our contention that ratings *directly* assessing the substantive content of jobs and tasks can contribute significantly to the understanding of work and to the practical usefulness of job analysis.

### Reliability of FJA Ratings

FJA is a method that emphasizes the controlled use of job language in analyzing task requirements (Cronshaw, 2012; Fine, 1955; Fine & Cronshaw, 1999). Within the context of the larger FJA system, FJA ratings have two major purposes: (a) to ensure that the task statements generated by focus groups are properly standardized by being written within the stringent guidelines required by the FJA theory and methodology; (b) to assist in the development of human resource interventions. The reliability of the FJA scales establishes that the content of the job task, standardized through the use of controlled job language, is sufficiently rigorous to ensure the consistency of job language within and between FJA tasks.

The purpose of this study was to assess the reliability of the FJA scales under their recommended conditions of administration: Ratings of tasks are made individually on the written tasks contained in the FJA task bank by trained job analysts, after which these raters arrive at a single set of consensus ratings. Consensus ratings are an added check to ensure that the tasks are written in strict accordance with the controlled language required by FJA. As well, the ratings can be used in the development of applications that are based on the job analysis data (see Fine & Cronshaw, 1999). We expect to find that when raters discuss and reach consensus ratings, the consensus ratings will be more reliable than individual ratings because discussion and consensus should eliminate unique errors or idiosyncrasies that may be present in individual ratings. In fact, Fine and Getkate (1995) recommend the use of consensus rather than independent ratings for exactly this reason. This study was conducted to test this assertion that the interrater reliability of the FJA scales will be higher for single consensus pairs than for single rater. Implications of the findings for the validity of the FJA system are then discussed.

## METHOD

### Task Stimuli

The rating stimuli provided to raters in this study were 50 task statements taken from a compendium of task statements used to benchmark, or provide a frame-of-reference for, FJA scale ratings (Fine & Getkate, 1995). There are over 450 task statements in the compendium generated from 65 different jobs. To generate task statements used as stimuli in this study, one task statement was randomly selected from Fine and Getkate at each level of seven FJA scales (the three Functional Skill [FS] scales, the three General Educational Development [GED] scales, and the Worker Instructions [WI] scale) for a total of 43 task statements. An additional seven task statements were selected at random over all of the remaining task statements to bring the total number of tasks rated up to 50. This sampling strategy for the task statements had two advantages: (a) The task statements reflected the widest possible variability in task attributes, lessening range restriction problems encountered in the Schmitt and Fine (1983) study where no task statements were present reflecting the highest levels on the Things Functional Scale; and (b) a large number of different jobs were represented in the study database, increasing the generalizability of our results. The six raters were instructed not to consult or discuss the Fine and Getkate (1995) book.

### Description of FJA Rating Scales

FJA is composed of 10 scales (see Table 1): Six ordinal scales assess the complexity of task attributes (i.e., Functional Skills: *Things, Data,* and *People* (TDP); and General Educational Development: *Reasoning, Mathematical,* and *Language*); the Worker Instructions scale, which measures the mix of prescription and discretion required by the task; and three scales assess the *orientation* of the task to T, D, and P). All of the first seven FJA scales are anchored by detailed and theoretically-derived descriptions at all levels of the scales. The last three orientation scales (i.e., relative/proportional emphasis on T, D, and P)—although not anchored by level-specific descriptions—are made with reference to the corresponding scale of Functional Skill (i.e., the rater refers to the levels assigned to the Functional skill scales when assigning the proportion of the task oriented toward T, D, and P). Among the FJA quality control measures is the requirement that highly trained FJA raters justify their ratings by referring to the systematic and theoretically driven wording of the FJA task statement.

### Description of FJA Raters and Rating Procedure

Six trained FJA raters rated the 50 tasks using the FJA scales and level definitions from Appendix A of Fine and Cronshaw (1999). All six raters were trained in FJA theory and methodology in accordance with the guidelines present-

[1] In FJA, the *Things, Data,* and *People* functional scales are each individually rated for their level of complexity, as well as their relative (i.e., proportional) emphasis placed on each in the task statement.

ed in Fine and Cronshaw (1999), and all raters previously had used FJA in professional consultations with organizations. They did not have particular experience with or knowledge of the jobs from which the task statements were derived. FJA ratings do not require the raters to be involved with the original job analysis of the job because all the information necessary to reliably rate the task is embedded in the task statement itself. The raters were asked to provide their ratings for all 10 FJA scales under both individual and consensus rating conditions. The data were collected consecutively at two points in time: (a) The six raters were first asked to independently rate 50 task statements and submit the results to the researchers for entry into the study database; (b) the independent ratings were returned to individual raters who were randomly assigned to one of three rating pairs with the instructions that they were to discuss their respective independent ratings and arrive at a single set of consensus ratings. The same raters were used over the two rating conditions as is usual in FJA practice. This procedure resulted in six sets of independent ratings and three subsequent sets of consensus ratings over the 50 FJA tasks and ten FJA rating scales.

### Analysis of FJA Scale Reliability Under Independent and Consensus Conditions

The *reliability* of the FJA scale ratings across the ten scales were obtained by conducting a Generalizability (G), then a Decision (D) study of the ratings (Shavelson & Webb, 1991). A G-study is a psychometric investigation that evaluates the relative contributions of various sources of error in a given measurement procedure (see Scherbaum, Dickson, Larson, Bellenger & Yusko, 2018 for an overview of generalizability theory). A D-study applies the results of the G-study to optimize the procedure for specific uses. The object of measurement for the G-study was *task*, and the single facet of generalizability (i.e., source of error) was either *individual rater* or *rater consensus pair*. The G-study used a fully crossed design such that *task* was fully crossed with both *rater* and *rater pair* (i.e., two separate analyses) such that all raters or rater pairs rated the same 50 tasks. In the D-study, generalizability coefficients for raters across tasks were run for a single rater/rater pair, then for six raters/rater pairs to provide a comparison to the earlier Schmitt and Fine (1983) results and for future research. These generalizability coefficients were then interpreted as a measure of interrater reliability.

Generalizability analyses were run on the FJA scale ratings for each of the independent and consensus pair rating conditions using the GENOVA program developed by Crick and Brennan (1983). These generalizability analyses, as mentioned, had two aspects: (a) a G-study component; and (b) a D-study component (Shavelson & Webb, 1991). The G-study design was *t* x *r* (*t*ask crossed with *r*ater) in the independent rating condition and *t* x *d* (*t*ask crossed with rater pair or *d*yad) in the consensus rating condition. Raters were treated as random effects to assess the consistency of ratings made by comparable sets of raters not included in this study, either individually or as consensus pairs.

## RESULTS

### Reliability of FJA Scales

Table 1 reports the means for the FJA scales investigated in this study, as well as differences in means between individual and consensus ratings for each FJA scale. When compared to the range of possible ratings (far-right column of Table 1), scale rating differences between the individual and consensus conditions were negligible except for Data and People orientation ratings where the differences between individual and consensus conditions, although small, were greater than 5% (in FJA raters are required to provide orientation ratings in increments of 5%).

We predicted that the interrater reliabilities of the FJA scales would be higher for a single consensus pair than for a single rater. When the generalizability results reported in Table 2 for a single random rater are compared to a single random consensus pair across the 10 FJA scales (Table 3), the use of consensus pairs is found to produce no overall improvement in interrater reliabilities over independent judgments when considered across the 10 FJA scales. A Friedman two-way analysis of variance by ranks comparing the reliabilities of the 10 FJA scales across the independent and consensus judgments yields a statistically nonsignificant result ($\chi^2$= .90, 1*df*). Therefore, the expectation that consensus ratings of the substantive content of FJA tasks will have greater reliability than independent ratings of the same content is not supported.

## DISCUSSION

It is demonstrated here that the substantively based scales, developed as an integral part of FJA theory and methodology, have high levels of interrater reliability. The interrater reliabilities found in this study for all but one of the ten FJA scales were higher, sometimes considerably so, than the overall mean reliability for job analysis ratings of .59 reported by Voskuijl and van Sliedregt (2002), which is below the lowest reliability obtained here for any of the FJA rating scales excepting for the FJA ratings for Mathematical Development. Similarly, Dierdorff and Wilson (2003) reported a sample-size weighted mean reliability of .63 over 10 studies that used job analysts as raters, with an 80% credibility interval of [.55, .71]. By comparison, the mean reliability average of .75 for a single rater over the 10 FJA scales reported in this study (Table 3) exceeds 80% of the reliabilities that Dierdorff and Wilson reported for studies—like this one—in which tasks were rated by job ana-

## TABLE 1.
Means for, and Differences Between, FJA Scales Under Individual and Consensus Rating Conditions

| | Scale means for conditions | | | |
|---|---|---|---|---|
| Scale name | Independent | Consensus | Absolute mean differences | Number of scale levels |
| Things function | 1.74 | 1.60 | 0.14 | 1-4 |
| Data function | 3.03 | 3.04 | 0.01 | 1-6 |
| People function | 2.55 | 2.36 | 0.19 | 1-8 |
| Things orientation | 30.90 | 28.95 | 1.95 | 5-100% |
| Data orientation | 38.63 | 45.95 | 7.32 | 5-100% |
| People orientation | 29.88 | 24.45 | 5.43 | 5-100% |
| Worker instructions | 3.62 | 3.39 | 0.23 | 1-8 |
| Reasoning development | 3.46 | 3.28 | 0.16 | 1-6 |
| Mathematical development | 2.05 | 2.02 | 0.03 | 1-5 |
| Language development | 3.28 | 2.97 | 0.31 | 1-6 |

## TABLE 2.
Variance Components for Ten FJA Scales Under Independent and Consensus Rating Conditions

| | Variance components for independent ratings (*N* = 6) | | | Variance components for consensus ratings (*N* = 3) | | |
|---|---|---|---|---|---|---|
| FJA | $\sigma^2_t$ | $\sigma^2_r$ | $\sigma^2_{t\,x\,r}$ | $\Sigma^2_t$ | $\sigma^2_d$ | $\sigma^2_{t\,x\,d}$ |
| Things function | .66<br>(.14) | .01<br>(.01) | .29<br>(.03) | .71<br>(.16) | .01<br>(.01) | .27<br>(.04) |
| Data function | 2.15<br>(.45) | .14<br>(.08) | .61<br>(.05) | 2.50<br>(.52) | .07<br>(.05) | .40<br>(.06) |
| People function | 1.92<br>(.41) | .11<br>(.07) | .91<br>(.08) | 2.05<br>(.46) | .06<br>(.05) | .68<br>(.10) |
| Worker instructions | 4.45<br>(.91) | .07<br>(.05) | .71<br>(.06) | 4.81<br>(.99) | .17<br>(.13) | .55<br>(.08) |
| Reasoning development | 2.01<br>(.41) | .03<br>(.02) | .35<br>(.03) | 2.17<br>(.45) | .07<br>(.05) | .29<br>(.04) |
| Math development | .90<br>(.20) | .08<br>(.05) | .62<br>(.06) | 1.29<br>(.28) | .02<br>(.02) | .40<br>(.06) |
| Language development | 1.94<br>(.40) | .02<br>(.01) | .37<br>(.03) | 1.93<br>(.40) | .03<br>(.02) | .30<br>(.04) |
| Things orientation | 823.52<br>(169.29) | 7.02<br>(5.26) | 137.16<br>(12.47) | 795.16<br>(164.28) | 3.19<br>(3.05) | 53.58<br>(7.73) |
| Data orientation | 461.48<br>(98.14) | 19.75<br>(12.46) | 174.20<br>(15.84) | 535.86<br>(114.25) | 10.31<br>(8.59) | 87.78<br>(12.67) |
| People orientation | 544.80<br>(112.79) | 1.56<br>(2.09) | 114.44<br>(10.40) | 560.83<br>(117.43) | 1.04<br>(1.64) | 60.77<br>(8.77) |

*Note.* In the variance components $t$ = task, $r$ = rater. Standard errors are presented in parentheses following the variance components. The large discrepancies between the magnitudes of the variance components of the orientation and other ratings are due in part to differences in the number of scale levels. For independent ratings, $df_t = 49$; $df_r = 5$; $df_{txr} = 245$. For consensus ratings, $df_t = 49$; $df_d = 2$, $df_{txd} = 98$.

TABLE 3.
Comparative Results for Reliability of Independent Raters and Consensus Pairs

| | Reliabilities for: | |
|---|---|---|
| FJA Rating Scale | One Independent Rater | One Consensus Pair |
| Things Function | .69 | .66 |
| Data Function | .74 | .74 |
| People Function | .65 | .79 |
| Worker Instructions | .83 | .81 |
| Reasoning Development | .83 | .86 |
| Math Development | .58 | .61 |
| Language Development | .83 | .78 |
| Things Orientation | .82 | .86 |
| Data Orientation | .70 | .82 |
| People Orientation | .82 | .83 |
| *Mean* | .75 | .78 |

lysts. The higher reliabilities we observe here are likely due to the theoretical grounding of and extensive development invested in the substantively based FJA scales.

Contrary to our expectations and the assertion by Fine and Getkate (1995), this study found that consensus pairs of experienced FJA raters produced no improvement in reliability over independent raters. This finding invites an explanation and an exploration of its relevance to other assessments where both individual and consensus ratings are made. Highhouse and Nolan (2012), in a historically based review of the evolution of assessment center theory and practice, note that a statistical combination of assessment center ratings will produce higher predictive validities (which is also a result found by Dilchert & Ones, 2009) and cost savings compared to a subjective combination of assessment center ratings made through consensus discussion. The superiority of statistically combining ratings over consensus discussion is likely due to the avoidance of various types of rating errors (e.g., halo, central tendency) that can occur when ratings are combined subjectively through consensus discussion. By extension, it is important to determine if consensus ratings of job and work analysis data would run into some of the same problems. Nevertheless, this study does suggest that consensus ratings work well for the substantive content approach. Also, it is important to remember that, in FJA, the unit of analysis is the work itself rather than the person performing the work. FJA scales usually remain separate for purposes of personnel decision making, unlike assessment center ratings, which are often combined in order to make a final accept/reject decision on a candidate for a job or promotion. FJA raters compare the content of task statements with detailed, substantively based descriptions in the FJA scales; and because FJA ratings are not combined, they are not subject to the problems associated with combining information into an overall subjective impression in the ensuing consensus discussion. Although further research is warranted, it appears that this decompositional approach of FJA ratings (whether individual or consensus) offers some of the same advantages—in terms of reduced bias and increased job relatedness—as the decomposed ratings that are used in statistical combinations of assessment center data. In short, it is likely that the extensive information in the FJA task statements and accompanying rating scales, when used together within a rigorous regime of rater training, allow even a single rater to achieve similar results to consensus ratings and these ratings are reliable enough to support their use in many, if not all, FJA-based HRM applications.

Differences in mean ratings between the individual and consensus conditions were negligible for most of the FJA scales; although small differences were found for the Data and People orientation scales where some care might need to be taken when using these scales in FJA projects. Anecdotal evidence from the raters in this research and other FJA projects suggests that raters find the three FJA orientation scales to be more ambiguous during the rating task than the other seven FJA scales, which all have detailed descriptions anchoring each level of the scale.

D-studies conducted across the FJA rating scales in both the independent and consensus conditions suggest that, taking into account a balance between maximizing reliabilities and containing costs, the optimal FJA rating strategy across the 10 scales is to use three independent raters (although, using a single rater yields acceptable results for many purposes). With the three-rater strategy, FJA reliabilities range from .74 - .91 with a mean of .85, which represents a significant improvement in practical terms over the reliability of consensus ratings reported in Table 3. When reporting ratings on individual tasks for purposes of quality control and FJA applications, the modal value across the three raters should be used (Fine & Cronshaw, 1999).

In general, the lesson for job and work analysis is that there is much to be gained in assessment rigor by providing very detailed task information in both task statements and the rating scales themselves, in addition to the rigorous training of job analysts. Perhaps improvements can be made to the extrapolated correlates approach in the psychometric adequacy of their ratings, including improved reliability, by the better anchoring of rating scales and the provision of more detailed and systematically written task information as stimuli for the rating task. We believe that such attempts would be well worth the effort.

It is a maxim in personnel assessment that "the reliability of a scale limits its validity." It is to this important question of the validity of job and work analysis data that we now turn. McCormick (1979) opined in his classic work on job analysis that "it is usually necessary to infer the validity of such [job analysis] data from evidence of their reliability as based on results from two or more independent analysts" (p. 34). In other sources, the validity of job analysis systems is equated with their ability to resource personnel interventions and decisions, that is, consequential validity (Sanchez & Levine, 2000): providing evidence of the content validity of achievement tests (Goldstein, Zedeck, & Schneider, 1993), validating employment tests via synthetic validity (McCormick, Jeanneret, & Mecham, 1972; Mossholder & Arvey, 1984; Primoff, 1959), and judging pay levels in job evaluation (Smith & Hakel, 1979). The job analysis system, as regards validity in its own right, largely remains a black box. Cronshaw, Best, Zugec, Warner, Hysong, and Pugh (2007) aimed to address this by developing a means to directly validate both qualitative and quantitative job analysis data. Their validation approach, developed with FJA in mind, proposed five interlocking strategies to ensure that FJA data is: (a) written in a standardized format (linguistic validation); (b) accurately reflects the experiences of the job incumbents (experiential validation); (c) generalizes across units and organizations (ecological validation); (d) makes theory-generated predictions that are confirmed by empirical research (hypothetico-criterial validation); and (e) supports organizational decision makers (social-organizational validation). This is the most wide-ranging model of job analysis validation reported in the literature. Although high levels of interrater reliability are assumed by all five FJA validation strategies, linguistic and experiential validation of FJA data are not possible without the high levels of interrater reliability demonstrated here. The other FJA validation strategies discussed by Cronshaw et al. (2007) are also informed by the high levels of interrater reliability of the FJA scales although these validation strategies require additional information collected in parallel with the FJA-guided collection of job data.

Two limitations present themselves in this study. First, the independent and consensus rating conditions built into this study were not independent. A traditional research design would require that the raters be randomly assigned across the two conditions from a larger population of raters, and, as a result, different raters would be used in the independent and consensus conditions. This latter approach was not taken because, in practice, FJA raters first make individual and independent judgments on task attributes and then meet to discuss their individually based ratings to arrive at a single consensus judgment. To reflect this practical reality in this study the same raters made both independent and consensus ratings, in that order. Nevertheless, a future study using random assignment across individual and consensus rating conditions could control for anchoring and adjustment effects that might yoke the ratings made in the consensus condition to the previous ratings made by the same raters in the independent condition.[2] Another potential concern is that the raters were not involved in collecting the task analysis data. This concern is based on how ratings are made in nonsubstantive content job analysis systems and is a misunderstanding of how FJA was developed and how it is used. Because FJA is a substantively based approach to job analysis, all the job information needed for reliable ratings is gathered and standardized in the task statements *before* the scale ratings are made. No prior knowledge or experience with the target job is necessary to inform ratings because all the information needed should be contained in the task statement. The need for supplemental information to make ratings is an indicator of an inadequately written task statement. Nevertheless, to rate FJA tasks, raters must be thoroughly trained and experienced in both the theory and methodology of FJA.

The results of this study provide good and bad news for job analysis generally. The good news is that reliable and consistent task ratings of generalized work activities can be obtained as long as: (a) the rating scales are anchored at all levels using theoretically-based definitions of generalized work attributes; (b) the task statements are written for the specific job and context using clear, descriptive language to ensure the accuracy, comprehensiveness, and usability of these narrative descriptions (Cronshaw et al., 2007); and (c) the raters are thoroughly trained in both the theory and methodology of the job analysis method. When it comes to FJA scale ratings, a single rater will usually suffice and consensus ratings by two or more job analysts are not needed to attain sufficiently high levels of reliability. The bad news is that based on the available research, only one method of job analysis presently available—FJA—meets the stringent conditions needed to produce such high reliability and consistency for single-rater, task-based ratings. We echo Landy and Farr's (1980)—among several others—call for greater

[2] We thank an anonymous reviewer for this insight.

emphasis on theoretically grounded work analysis research (e.g., Morgeson & Campion, 1997; Morgeson, Spitzmuller, Garza, & Campion, 2016). We believe that a sustained effort in theory development and the rigorous control of the job language would improve the reliability, validity, and utility of job and work analysis research and practice.

## REFERENCES

Crick, J. E., & Brennan, R .L. (1983). A generalized analysis of variance system (Version 2.1) [Computer software]. Iowa City, IA: The American College Testing Program.

Cronshaw, S.F. (2012). Functional Job Analysis (Chapter 15). In M. A. Wilson, W. Bennett, S. G. Gibson, & G. Alliger (Eds.), The handbook of work analysis in organizations: Methods, systems, applications, and science of work measurement in organizations. New York, NY: Routledge/Psychology Press.

Cronshaw, S. F., Best, R., Zugec, L., Warner, M. A., Hysong, S. J., & Pugh, J. A. (2007). A five-component validation model for functional job analysis as used in job redesign. Ergometrika, 4, 12-31.

Dierdorff, E. C., & Wilson, M.A. (2003). A meta-analysis of job analysis reliability. Journal of Applied Psychology, 88, 635-646.

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. International Journal of Selection and Assessment, 17(3), 254-270.

Fine, S. A. (1955). Functional job analysis. Journal of Personnel Administration & Industrial Relations, 2, 1-16.

Fine, S.A., & Cronshaw, S. F. (1999). Functional job analysis: A foundation for human resources management. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fine, S. A. & Getkate, M. (1995). Benchmark tasks for job analysis: A guide for functional job analysis (FJA) scales. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W.C. Borman (Eds.), Personnel selection in organizations (pp. 3-34.). San Francisco, CA: Jossey-Bass.

Highhouse, S., & Nolan, K. P. (2012). One history of the assessment center (pp. 25-44). In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.). The psychology of assessment centers. New York, NY: Routledge.

Landy, F. J. & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

McCormick, E. J. (1979). Job analysis: Methods and applications. New York, NY: AMACOM.

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(4), 347-368.

Morgeson, F. P. & Campion, M.A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. Journal of Applied Psychology, 82(5), 627 655.

Morgeson, F. P., Spitzmuller, M., Garza, A. S., & Campion, M. A. (2016). Pay attention! The liabilities of respondent experience and carelessness when making job analysis judgments. Journal of Management, 42(7), 1904-1933.

Mossholder, K. W., & Arvey, R. D. (1984). Synthetic validity: A conceptual and comparative review. Journal of Applied Psychology, 69(2), 322-333.

Primoff, E. S. (1959). 4. Empirical validations of the J-coefficient. Personnel Psychology, 12(3), 413-418.

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

Scherbaum, C., Dickson, M., Larson, E., Bellenger, B., Yusko, K., & Goldstein, H. (2018). Creating test score bands for assessments involving ratings using a generalizability theory approach to reliability estimation. Personnel Assessment and Decisions, 4(1), 1-8.

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.

Shavelson, R. J. & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

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(4), 677-692.

Voskuijl, O.F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. European Journal of Psychological Assessment, 18(1), 52-62.

RECEIVED 09/29/17 ACCEPTED 01/23/19