Assessing and Reconciling Between-Group Differences in Job Analysis Ratings

Matthew Castillo
Louisiana Tech University

Christina Cantu
Louisiana Tech University

Frank P. Igou
Louisiana Tech University

Job analysis is an important tool underlying several organizational functions including performance appraisals, selection, and training. Typical job analyses are conducted with a random, stratified sample of subject matter experts (SMEs). These individuals are often job incumbents who are asked to rate and make judgments concerning the frequency and importance of job tasks. It is presumed that incumbents who perform a job should evaluate it similarly. Accordingly, it may be concerning if there is significant variance in employees' job analysis ratings depending on what is causing the differences. As such, job analysts must determine if rating differences are legitimate ("real") or due to error. This review provides an overview of various statistical techniques that can be used to assess the significance of between-group differences in job analysis ratings and how they can be used by job analysis experts to determine the extent to which SME ratings are consistent. In addition to detecting rating differences among SME groups, we also provide practical recommendations for reconciling those differences, if necessary.

Keywords: job analysis, rating differences, reconciliation sessions

INTRODUCTION

Job analysis is an important tool underlying several organizational functions including performance appraisals, selection, and training (Weekley et al., 2019). In fact, it has been described as the fundamental building block for all employment decisions (Cascio & Aguinis, 2005). Job analysis provides organizations with information about the relative importance of essential work behaviors, tasks, and job requirements (Gael, 1988). Typical job analyses are often conducted with a random, stratified sample of subject matter experts (SMEs). Although there are no universally accepted criteria for who may be considered an SME, these individuals are often job incumbents who are asked to rate and make judgments concerning the frequency and importance of job tasks (Truxillo et al., 2004). Supervisors and retired incumbents (referred to as global SMEs) are also sometimes used. The combination job analysis method (C-JAM; Levine, 1983) is a common job analysis method that identifies work characteristics (e.g., tasks and work behaviors) and

worker characteristics (knowledges, skills, abilities and other characteristics; henceforth "KSAOs") and links them by asking SMEs to rate the relative importance of each KSAO to each task or work behavior.

Although it is possible for specific job activities to vary and people working under the same job title to perform very different tasks, it is presumed that incumbents who perform a job similarly should also evaluate it similarly. Accordingly, it may be concerning if there is significant variance in employees' job analysis ratings depending on what is causing the differences (Conte et al., 2005). Investigating differences in job analysis ratings can help determine the transportability of job analysis information, or how similar job information is across organizations (Li et al., 2008). This is important to consider if applying job information from a generic job database, such as the Occupational Information Network (O*NET; Peterson et al., 2001). As such, job analysts must determine if rating differences are legitimate ("real") or due to error. Too often, job analysis data is not examined but taken at face value. There is little research examining job analysis ratings although similar processes have been used in research and practice on other rating systems such as performance appraisals (Igou et al., 2019). Various statistical techniques can be used to assess the significance of between-group differences in job analysis ratings. This review adds to extant literature by providing an overview of those techniques and how they can be used by job analysis experts to determine the extent to which SME ratings are consistent. In addition to detecting rating differences among SME groups, we also provide practical recommendations for reconciling those differences if necessary.

SOURCES OF RATING DIFFERENCES

Results from job analyses are often used to infer the KSAOs required to successfully perform tasks and work duties (Brannick. Levine, & Morgeson, 2020). This information is the legal foundation for human resource interventions (Morgeson & Campion, 1997). For example, job analyses reveal essential job functions which can be used to guide reasonable accommodations for disabled applicants (Richman & Quiñones, 1996). Legal guidelines following court cases regarding selection and employment decisions based on job analyses increased interest in factors that influence outcomes of job analysis procedures (Mullins & Kimbrough, 1988). Previous literature has acknowledged the scarce investigation of reliability issues in job analysis (Harvey, 1991). There is a lack of consensus regarding variables that affect job analysis reliability and validity (Dierdorff & Wilson, 2003). Subsequent research has called for a deeper understanding of disagreement within jobs (Morgeson & Campion, 2000). Harvey (1991) notes that human judgment is subject to inaccuracies which could be influenced by systematic or unconscious factors. Some examples include SME job experience and personal characteristics such as demographic, social, and cognitive variables as well as work attitudes (Conte et al., 2005; Morgeson & Campion, 1997; Tross & Maurer, 2000). Other errors could result from individuals trying to move through the rating process too quickly, biases, fatigue, and distractibility. A meta-analysis of job analysis reliability from 46 studies revealed that job incumbents had lower reliabilities than job analysts or technical experts (Dierdorff & Wilson, 2003). However, adding specific task data improves reliability compared to general activity statements. Further, rater training such as frame-of-reference training that provides a category system for ratings can improve reliability compared to no training (Sanchez & Levine, 2009).

Factors that influence rating differences may be categorized as legitimate, or "real job", differences and error sources (Morgeson & Campion, 1997). Legitimate differences in how a job is performed provide potentially useful job-related information (Sanchez & Levine, 2009). SMEs may have differences in how they perceive or perform a job. Tross and Maurer (2000) suggested that longer tenured or more experienced SMEs may be more likely to assign higher ratings of importance to tasks or KSAOs. Similarly, it has been suggested that incumbents who are high performers, highly motivated, or high on organizational commitment may provide inflated job-analysis ratings (Borman, Dorsey, & Ackerman 1992; Conte et al., 2005). Moreover, the problems found in other situations using SME ratings are likely to be found in job-analysis ratings, such as conformity pressures, lack of motivation, social desirability of response, information overload or deficiency, carelessness, and method effects (see Morgeson & Campion, 1997 for more information). Organizations concerned about workforce inclusivity and Equal Employment

Opportunity (EEO) issues may want to examine rating differences by demographic groups. To reduce the likelihood that selection and promotional procedures may have adverse impact and ensure adherence to the Uniform Guidelines (EEOC, 1978), it may be necessary to include members of EEOC protected groups as SMEs on both the job analysis and test development processes due to the potential for liabilities.

ASSESSING RATING DIFFERENCES

Differences in SME ratings could have a major influence on reliability if they are due to perceptual difference or demographic characteristics as opposed to real differences in how jobs are performed. Information concerning the amount of error present in job analysis data reveals how much confidence can be placed in it and whether decisions should be made based on it (Brannick et al., 2020). Thus, it is important to determine if rating differences are real or due to uncontrolled measurement error. Differences due to information specific to the way an individual experiences their job has been referred to as idiosyncratic variance (Morgeson & Dierdorff, 2011). The traditional approach of aggregating ratings could oversimplify responses and neglect idiosyncratic variance. This review will focus on identifying and evaluating idiosyncratic variance via various statistical techniques. Prien et al. (2003) suggests adopting a practical model of evaluating accuracy of job analysis ratings in a series of steps. These include sorting data into groups, evaluating rating consistency, and comparing group means. Several statistical techniques can be used to carry out the above steps and assess between group differences in job analysis ratings. Specifically, intraclass correlation (ICC), Fisher's *r* to *z* transformation, t-test, and analysis of variance (ANOVA) will be discussed.

Intraclass Correlations

Interrater agreement is a measure of the degree of similarity in ratings often expressed in terms of percentage of agreement, or a within-group correlation (Morgeson & Campion, 2000). Intraclass correlation coefficients (ICCs) are a statistical measure of rater agreement ranging from 0 (no agreement) to 1 (perfect agreement; Leeds & Griffith, 2001). It is a score of the degree of consensus in ratings given by various judges. They provide the expected reliability for a single judge's rating and are commonly used to examine intra- and interrater reliability (Koo & Li, 2016; Shrout and Fleiss, 1979). ICCs may be the best choice for examining consistency between raters because they center and scale data using a pooled mean and standard deviation, unlike correlation coefficients which center each variable by its own mean and standard deviation. Further, ICCs are not limited to a meaningful pairing between only two raters like the Pearson correlation. There are six versions of ICC depending on the level of analysis (single vs average measures), type of ANOVA (one vs two way), and effect of raters (fixed vs random). For job analyses, ICC (2, k) is likely the most appropriate statistical procedure to use. This iteration is referred to as a two-way random, average score ICC and measures absolute agreement rather than consistency (McGraw & Wong, 1996). Agreement percentages tend to inflate reliability (James, Demaree, & Wolf, 1984). Thus, ICCs are a better way to estimate interrater agreement via the average deviation index which provides a direct estimate of agreement (Dunlap, Jones, & Bittner, 1983).

Fisher's *r* to *z*

The Fisher's *r* to *z* transformation converts r statistics to standard z scores to determine whether there are statistically significant differences between two correlation coefficients (r_a and r_b). It can also be used to examine differences in ICC's between groups. Applied to job analysis, this statistical technique allows comparison of between-group differences in rating processes (Surrette, Aamodt, & Johnson, 1990). For example, are job analysis ratings by incumbent SMEs or global SMEs more consistent than the other or are they statistically the same? Since job analyses often rely on small to moderate sample sizes of SMEs, it is important to consider how best to improve reliability while maintaining the underlying relationships. One approach is to average ratings before calculating the statistic of interest. A second approach is to find correlations between all ratings then average the correlations in blocks. The latter produces a skewed sampling distribution when the population parameter is significantly different than zero. This results in a

biased average *r* which underestimates the actual population correlation (Kendall & Stuart, 1979). To correct skew, or mean or modal shifts, correlations can be converted to standard scores via Fisher's *r* to *z* transformation prior to averaging. Once averaged, tests for comparisons of means can be used to compare averages and look for between group rating differences. These include t-tests and analyses of variance (ANOVAs).

T-Test

T-tests can be used to compare two groups from a sample regarding whether a correlation coefficient is significantly different from zero (Field, 2009). There are two types of t-tests that could be used to compare SME ratings: the one sample or independent samples t-test. A one sample t-test would be appropriate to check if the sample mean is different from a hypothesized value (usually zero; Park, 2009). This could be used if the SMEs selected to provide job analysis ratings all came from the same population (i.e., job incumbents at the same level in an organization with the same title). Between, or independent, group differences require an independent samples t-test. This would be appropriate if SMEs selected to provide job analysis ratings came from different populations (i.e., job incumbents and their supervisors). To use this statistic in job analyses, mean ratings of task statements and KSAOs must be calculated. Then, the sample of SMEs must be divided into two groups. To meet the t-test assumption of random sampling, SMEs should be randomly assigned to their group. However, in the applied world assignment may not be truly random. It could be argued that practitioners try to select a representative sample but end up with a convenience sample due to practical limitations. Finally, the t-test can be used to compare the mean ratings of the two groups to look for significant differences. Comparisons are made per condition rather than comparing difference groups. The *t* statistic is the mean difference between two groups divided by the standard error of measurement and thus indicates the meaningfulness of any difference. The standard error is used to gauge the variability between means where smaller numbers indicate similar means (Field, 2009). The null hypothesis states that the means of the two groups are equal, therefore the difference between the two means should be zero. Failing to reject the null hypothesis would indicate that SME ratings are not significantly different from one another. To determine if a t-test is significant, the observed value is compared to a critical value. If the observed *t* is larger than the critical value, the null hypothesis is rejected indicating significant mean differences. One may also consider using an unadjusted t-test. Although this inflates the Type I error rate, the purpose of comparing job analysis ratings between groups is not to support a hypothesis or theory but rather as a red flag that there may be differences indicating the potential need to follow up with SMEs using qualitative processes to determine whether there are real difference that need to be addressed in terms of job analysis conclusions or differences that have no practical significance despite statistical significance. Although t-tests (as well as ANOVAs, described below) are fairly robust with respect to the violation of assumptions, data that violate t-test assumptions such as normality usually require non-parametric tests including the Wilcoxon Rank Sum and Mann-Whitney U tests (see Blair & Higgins, 1980 and Sawilowsky, 2005 for more info).

Analysis of Variance

Analysis of variance (ANOVA) procedures can be used to assess dyadic or group differences (DeCoster, 2002). For job analyses, ANOVAs allow analysts to compare average aggregate ratings of tasks and KSAOs by SME subgroups to determine if there are significant differences. Similar to t-tests, ANOVAs compare group means. They are an extension of the t-test in that they allow mean comparisons for more than two groups. Instead of the *t* statistic, ANOVAs use the *F* statistics to test if three or more groups have the same mean. Again, the observed *F* is compared to a critical value and if the observed value is greater than the critical value the null hypothesis is rejected indicating significant mean differences. The rationale of why to use ANOVA rather than conducting multiple t-tests is that the probability of falsely rejecting the null hypothesis (Type I error) increases with each t-test (Field, 2009). Inflating error rates decreases the probability that a significant mean difference reflects a true difference rather than a difference expected due to chance. ANOVA is considered a robust test in that it controls for Type I error even when data are non-normal. An example of this is that rating distributions can be left skewed if raters agree about the importance

of tasks and KSAOs. In applied settings, it may be acceptable to inflate the probability of Type I error if there are possible discrepancies that need to be locally discussed and resolved compared to researchers presenting conclusions that could lead science in the wrong direction. It is important to note that sample sizes should still be equal to conduct ANOVAs. Violating this assumption may call for the use of non-parametric tests, such as the Kruskall-Wallis test (see Feir-Walsh & Toothaker, 1974 and Hecke, 2012 for more info). However, given that ANOVA is robust with respect to violation of assumptions it may be ok to still use it for the purpose of detecting between-group differences in job analysis ratings although it increases the likelihood of yielding a Type II error.

Unlike t-tests, a significant *F* value does not tell you which group means are significantly different from one another, only that a significant difference exists. With t-tests, a significant *t* would tell you that group one had a significantly difference mean than group two. Since there are more than two groups in an ANOVA, post hoc analyses must be conducted after finding a significant *F* value to determine which groups differ. This is accomplished via pairwise comparisons between all groups using t-tests. There are several different post-hoc procedures depending on the situation, but two common tests to use in job analyses are the Bonferroni correction and Tukey's honestly significant difference (HSD) because they control for Type I error rates (Field, 2009). The Scheffe procedure could also be used seeing as how it is most appropriate when the purpose of analysis is to explore the data (Ruxton & Beauchamp, 2008). However, the procedure is one of the most conservative post hoc tests and is thus more likely to conclude there is no difference when one may exist (Type II error). The type of post hoc test used is important because some tend to be too conservative which could result in a loss of power meaning differences could be rejected that are truly meaningful.

RECONCILING RATING DIFFERENCES

Job analysis data can be examined to determine whether there may be meaningful job differences in jobs using the same title or whether SMEs are actively participating and attending to the same aspect of the job. Detecting true rating differences via statistical techniques is important because it may not be possible to easily aggregate data for certain jobs and develop generic descriptors and requirements needed to create job descriptions or develop minimum qualifications (MQs). Rating differences that are not due to the above could result from rater error such as perceptual differences in the frequency or importance of difference aspects of a job or biases such as self-presentation bias (Buckley et al., 2007). Rating differences may be discussed in reconciliation sessions once initial ratings are made by SMEs. For example, a job analysis questionnaire (JAQ) could be completed by SMEs using networked laptops with all SMEs while a job analyst or similarly trained professional monitors SME's ratings in real time. Working in real time with live groups or meeting with SMEs as they are reviewing results of the rating process provides the chance to reconcile discrepant ratings in terms of perceptions of the job or real differences in how individuals perform jobs. These sessions provide the means for SMEs to explain why they gave a particular rating including what factors were considered when making their decision. Reconciliation sessions allow SMEs to change their initial rating which could be influenced by other individual's perceptions leading to bias. Previous research asserts that variance in job analysis ratings could be due to demographic variables, levels of performance, social and cognitive variables, and work attitudes (Li et al., 2008). As such, practical guidance for addressing these types of situations would benefit organizations and practitioners when using SMEs to provide job analysis ratings to ensure reconciliation sessions are developed and administered in a structured, bias-free manner. Reconciliation sessions during the job analysis process can be compared to discussion meetings, calibration sessions, and the process of meeting during content analysis to discuss ratings. Although a formal process is not outlined in the job analysis literature, this review is intended to further establish guidelines for practitioners to understand the reconciliation process.

Reconciliation sessions have been informally described by practitioners as meetings in real-time, either as SMEs are performing ratings or after ratings have been assigned, to discuss and reconcile notable differences (Igou, Girardot, & Wright, 2019). During these types of reconciliations, SME's often discuss differences in assigned ratings to establish a common frame of reference (FoR) and calibrate their ratings

of different criteria. The resolution of any differences may be done on a judgmental or subjective basis without defining a threshold for meaningful difference. For example, some rating sessions may require that SME ratings be within one point of each other while others may view two points as acceptable. Speer, Tenbrink, and Schwendeman (2019) assert that there is little scholarly research on implementing calibration processes with searches yielding anecdotal accounts and generalized instructional guidelines. Although, to the best of the authors' knowledge, this process has not been outlined specifically in the job analysis literature it has been documented in other I-O related areas where SMEs are tasked with rating criteria and reconciling rating differences. For example, in selection a panel of judges commonly evaluates candidates during structured oral interviews (SOIs) and rates job applicants' performance in various areas according to pre-established benchmarks. While there does not appear to be a standardized method for reconciling rater differences, descriptions of this process can be found within structured oral interview (SOI), performance management, and qualitative research literature, which describe group discussions (Buckley et al., 2007; Van Iddekinge et al., 2006), calibration meetings (Isaacs et al., 2020; Speer et al., 2019), and inter-rater reliability (IRR) reconciliation (McDonald et al., 2019). A review of this literature reveals that, regardless of the method, reconciliation sessions appear to reduce variance in SME ratings which increase accuracy and IRR (McDonald et al., 2019; Roch, 2006; Speer at al., 2019).

As mentioned previously, during rating sessions assessors often make independent ratings before having a discussion session about their ratings. The discussion session may or may not require consensus among raters (i.e., absolute agreement; Buckley et al., 2007; Van Iddekinge et al., 2006). When not using a consensus, raters may use the discussion session to discuss ratings that differ to a pre-established amount (i.e., by more than two points) and talk about their ratings until they agree (i.e., to within two points; Van Iddekinge et al., 2006). Depending on the procedure or situation, the discussion session may be followed by an additional rating session where an SME changes their initial rating to be within the established rating criteria (Buckley et al., 2007). For example, if SME A gave an initial rating of 2 and SME B gave an initial rating of 5 on some job analysis dimension (i.e., importance or critically of a particular job task or KSA) then they would need to discuss why they assigned that particular score. After the reconciliation session, SME A would either need to bring their rating up to a 3 or SME B would need to bring their rating down to a 4 in a follow-up rating session to be within the established rating criteria of two points. According to social identity and similarity-attraction theories, SME ratings can be influenced by who the raters are, whom (or what) they are rating, and with whom they are rating (Buckley et al., 2007). Thus, it is important for managers and organizations to consider all three factors when forming rating panels to ensure SME ratings are free from bias. Including raters with different perspectives can lead to better coverage of the criterion domain which increases total ratee information leading to improved rating accuracy (Speer et al., 2019). Roch (2006) provides examples of previous interventions to improve rating quality including providing rewards for accurate ratings, discussing ratings with ratees, and scrutiny by an expert.

Prior to completing a rating task, SMEs are often required to attend a rater training session. Calibration sessions are a type of rater training that is used so that raters establish a common frame of reference by analyzing ratings from other raters on the same dimensions so that they understand differences in ratings according to benchmarks or other criteria (Gunnell, Fowler, & Colaizzi, 2016; Isaacs et al., 2020). The calibration process has been defined as time-bound efforts where criteria is collectively discussed by a group of raters prior to making final ratings (Speer at al., 2019). It is typically broken down into three stages: pre-evaluation, rating calibration meeting, and final ratings. During pre-evaluation, raters must have pre-defined notions of how they intend to rate the criteria of the rating task. Initial ratings are considered "pre" ratings because they are not official but are used as prior evaluations to guide discussions during the next stage. Second, raters must formally meet with others to discuss their ratings. This process can vary from organization to organization but should involve collective discussions among raters wherein raters justify their ratings and refine collective judgments prior to final ratings (Pytel & Hunt, 2017). Finally, once raters have met to discuss their ratings, they must make their final ratings (Speer et al., 2019). This can be done during the rating calibration meeting itself or shortly after the meeting. Benefits of rating calibration meetings include improving inter-rater reliability (IRR) and reducing variance attributable to raters.

The key to rating calibration meetings to reconcile rating differences is that there should be clearly defined rating criteria, such as with rubrics or behaviorally anchored rating scales (BARS), and raters should demonstrate alignment with established, predefined benchmarks (Isaacs et al., 2020). These common conceptualizations are similar to frame-of-reference (FOR) training which leads to a shared meaning of various criteria and application of common standards to ratings. Further, measures should be taken to ensure that raters are made accountable for their ratings by justifying them which can result in more thoughtful effort in making initial evaluations and higher rating accuracy (Speer et al., 2019). Finally, managers and organizations need to be aware of the perceived motives behind calibration meetings to ensure that ratings are not distorted to achieve alternative goals such as artificially suppressing or inflating ratings for various reasons. As such, the purpose and structure of calibration meetings should be discussed and explicitly stated prior to selecting rater panels and assigning SMEs a rating task. Often employees and SMEs are suspicious of consultants or job analysts coming into their work unit and asking questions making it likely they will perceive them as a threat and think the real reason is to make their work lives more difficult or downsize. However, most job analysis projects can be done with complete transparency. It is important to get buy-in for the project from the top down and for managers and supervisors to announce the job analysis activities well ahead of when they are scheduled. Further, management should be willing to address any questions and allay any concerns which could foster feelings of trust and psychological safety. Finally, SMEs should be informed that they are the key to the process and that their work could lead to lasting changes and improvements in the organization.

CONCLUSION

This review provided information regarding the use of simple statistical techniques for examining between-group differences in job analysis ratings as well as guidance for reconciling those differences. Considerations as to whether data fit test assumptions and what population SME samples are drawn from are important in choosing the appropriate statistical procedure. Intraclass correlation (ICC), Fisher's *r* to *z* transformation, t-test, and analysis of variance (ANOVA) were discussed in the context of assessing and identifying meaningful differences in average job analysis ratings on task statements and KSAOs. This information is important in ensuring best practice in conducting job analyses as well as determining if SMEs understand the nature of the job they are rating and agree on the required KSAOs and essential job tasks. When true rating differences exist between SME ratings, it is necessary to reconcile those differences via reconciliation sessions such as having SMEs discuss why they assigned a specific rating. Recommendations included clearly defining the rating criteria, training raters to ensure they align with the rating criteria, holding raters accountable for their ratings, ensuring there is no rater bias or social desirability effects on ratings, and explicitly stating the purpose and structure of rating calibration meetings.

REFERENCES

Blair, R.C., & Higgins, J.J. (1980). The power of t and Wilcoxon statistics. *Evaluation Review*, *4*, 645–656. doi:10.1177/0193841x8000400506

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.

Brannick, M.T., Levine, E.L., & Morgeson, F.P. (2020). *Job and work analysis: Methods, research, and applications for human resource management* (3rd ed.). Los Angeles, CA: Sage Publications.

Buckley, M.R., Jackson, K.A., Bolino, M.C., Veres, J.G., & Feild, H.S. (2007). The influence of relational demography on panel interview ratings: A field experiment. *Personnel Psychology*, *60*, 627–646. doi: 10.1111/j.1744-6570.2007.00086.x

Cascio, W.F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson Education.

Conte, J.M., Dean, M.A., Ringenbach, K.L., Moran, S.K., & Landy, F.J. (2005). The relationship between work attitudes and job analysis ratings: Do rating scale type and task discretion matter? *Human Performance*, *18*, 1–21.

DeCoster, J. (2002). *Using ANOVA to examine data from groups and dyads*. Retrieved from http://www.stat-help.com/notes.html

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

Dunlap, W.P., Jones, M.B., & Bittner, A.C. (1983). Average correlations vs. correlated correlations. *Bulletin of the Psychonomic Society*, *21*, 213-216.

Equal Opportunity Employment Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, *43*(166), 38290–38315. Washington, DC: Equal Employment Opportunity Commission.

Feir-Walsh, B.J., & Toothaker, L.E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, *34*, 789-799. doi: 10.1177/001316447403400406

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Gael, S. (1988). Job descriptions. In S. Gael (Ed.), *The Job Analysis Handbook for Business, Industry, and Government* (pp. 71-89), New York: Wiley.

Gunnell, K.L., Fowler, D., & Colaizzi, K. (2016). Inter-rater reliability calibration program: Critical components for competency-based education. *Competency-Based Education*, *1*, 36-41. doi: 10.1002/cbe2.1010

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

Hecke, T.V. (2012). Power study of anova versus Kruskal-Wallis test. *Journal of Statistics and Management Systems*, *15*, 241-247. doi: 10.1080/09720510.2012.10701623

Igou, F., Girardot, R.E., & Wright, M. (2019, April). *Assessing and reconciling between-group differences in job analysis ratings*. Symposium presented at the 34th Annual Society for Industrial and Organizational Psychology Conference, National Harbor, MD.

Isaacs, A.N., Miller, M.L., Hu, T., Johnson, B., & Weber, Z.A. (2020). Inter-rater reliability of web-based calibrated peer review within a pharmacy curriculum. *American Journal of Pharmaceutical Education*, *84*.

James, L.R., Demaree, R.G., & Wolf, G. (1993). Rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*, 306-309. doi: 10.1037/0021-9010.78.2.306

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics.* London: Charles Griffin.

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–164.

Landers, R.N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower*, *2*.

Leeds, J.P., & Griffith, R. (2001). Critical incident inter-rater agreement among security subject-matter experts. *Journal of Security Administration*, *24*, 31-44.

Levine, E.L. (1983). *Everything you always wanted to know about job analysis*. Tampa, FL: Mariner Publishing Company.

Li, W-D., Wang, Y-L., Taylor, P., Shi, K., & He, D. (2008). The influence of organizational culture on work-related personality requirement ratings: A multilevel analysis. *International Journal of Selection and Assessment*, *16*, 366-384.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1-23.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. doi: 10.1037/1082-989X.1.1.30

Morgeson, F.P., & Campion, M.A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, *82*, 627–655.
Morgeson, F.P., & Campion, M.A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior*, *21*, 819–827.
Morgeson, F.P., & Dierdorff, E.C. (2011). Work analysis: From technique to theory. *APA Handbook of Industrial and Organizational Psychology*, *2*, 3-41.
Mullins, W.C., & Kimbrough, W.W. (1988). Group composition as a determinant of job analysis outcomes. *Journal of Applied Psychology*, *73*, 657–664.
Park, H.M. (2009). *Comparing group means: T-tests and one-way ANOVA using STATA, SAS, R, and SPSS*. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.
Peterson, N.G., Mumford, M.D., Borman, W.C., Jeanneret, P.R., Fleishman, E.A., Levin, K.Y., . . . Dye, D.M. (2001). Understanding work using the Occupational Information Network (O*NET): Implications for practice and research. *Personnel Psychology*, *54*(2), 451–492.
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.
Pytel, L., & Hunt, S. (2017). *Total workforce performance management: Using talent calibration to effectively manage the reality that all employees are valuable, but some employees are more valuable than others*. Retrieved from https://www.slideshare.net/BhupeshChaurasia/using-calibration-effectively-total-workforce-performance-management
Richman, W.L., & Quiñones, M.A. (1996). Task frequency rating accuracy: The effect of task engagement and experience. *Journal of Applied Psychology*, *81*, 512.
Roch, S.G. (2006). Discussion and consensus in rater groups: Implications for behavioral and rating accuracy. *Human Performance*, *19*, 91–15. doi: 10.1207/s15327043hup1902_1
Ruxton, G.D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral Ecology*, *19*, 690-693. doi: 10.1093/beheco/arn020
Sanchez, J.I., & Levine, E.L. (2009). What is (or should be) the difference between competency modeling and traditional job analysis? *Human Resource Management Review*, *19*, 53-63.
Sawilowsky, S.S. (2005). Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney U test for shift in location parameter. *Journal of Modern Applied Statistical Methods*, *4*, 598-600.
Schmitt, N., & Fine, S.A. (1983). Inter-rater reliability of judgements of functional levels and skill requirements of jobs based on written task statements. *Journal of Occupational Psychology*, *56*, 121–127.
Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420.
Speer, A.B., Tenbrink, A.P., & Schwendeman, M.G. (2019). Let's talk it out: The effects of calibration meetings on performance ratings. *Human Performance*, *32*, 107-128. doi: 10.1080/08959285.2019.1609477
Surrette, M.A., Aamodt, M.G., & Johnson, D.L. (1990). Effects of analyst training and amount of available job related information on job analysis ratings. *Journal of Business and Psychology*, *4*, 439-451.
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.
Truxillo, D.M., Paronto, M.E., Collins, M., & Sulzer, J.L. (2004). Effects of subject matter expert viewpoint on job analysis results. *Public Personnel Management*, *33*, 33-46.
Van Iddekinge, C.H., Sager, C.E., Burnfield, J.L., & Heffner, T.S. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. *International Journal of Selection and Assessment*, *14*, 193–205. doi: 10.1111/J.1468-2389.2006.00352.X

Weekley, J.A., Labrador, J.R., Campion, M.A., & Frye, K. (2019). Job analysis ratings and criterion-related validity: Are they related and can validity be used as a measure of accuracy? *Journal of Occupational and Organizational Psychology*.
Wendler, C., Glazer, N., & Cline, F. (2019). Examining the calibration process for raters of the GRE general test. *ETS GRE Board Research Report*.