Using $r_{WG(J)}$ and the Intraclass Correlation Coefficient (ICC) as Measures of Rater Agreement and Reliability

Donald D. Conant, Ph.D.
Associate Professor of Business
Saint Martin’s University
5000 Abbey Way SE
Lacey, WA 98503
USA

Abstract

When measuring social constructs such as organizational culture and climate, researchers may aggregate perceptual measures without first establishing perceptual agreement. This results in an average of the ratings without first determining the extent to which the raters are experiencing, and thereby rating, the same social construct. Interrater agreement ($r_{WG(J)}$) and reliability (ICC) are common methods used to determine perceptual agreement. Students and researchers should be proficient in using and interpreting these measures. In this article I review one measure of interrater agreement and three measures of interrater reliability. For each I provide a brief history of their development, a justification for their usage, and selected research which serves as a benchmark for interpreting their results.

Keywords: interrater agreement, interrater reliability, organizational culture, organizational climate, perceptual agreement

1. Introduction and Rationale

In research where the research variables include organizational culture or climate, it is common for researchers to use a survey instrument to quantify the culture or climate variable. Selected participants each complete a survey after which the researcher aggregates the results. The aggregated results are then put forth as a measure of the culture or climate of the organization. However, before aggregation is justified some measure of perceptual agreement among the participants must be made. It is at this point that errors can occur in the selection of the appropriate measures of rater agreement and reliability.

The purpose of this article is to review two indices commonly used to justify the aggregation of data on organizational culture and climate; $r_{WG(J)}$ and the Intraclass Correlation Coefficient (ICC). Researchers using these indices should be aware of the history of their development, their various forms, research based benchmarks for interpreting results, and how statistical software can be used in the calculation of results.

1.1 Culture Measures and Data Aggregation

Organizational culture is a shared perception. To measure it, individual perceptions are combined in a way that reflects the culture as a whole. This combining is not intended to provide an average of disparate experiences. Rather, it should provide insight into the depth of perceptual agreement that exists regarding a particular culture. In studying organizational climate James (1982) concluded that

> a climate construct at the aggregate level is defined in precisely the same manner as it is at the individual level. For example, a shared perception that an environment is ambiguous allows one to describe that environment as ambiguous…[However] perceptual agreement must be demonstrated before climate scores are aggregated. (p. 221)

The ability to measure perceptual agreement (consensus) within or among groups plays an important role in research (Conway & Schaller, 1998). Two of the more common measures of this consensus are interrater agreement and interrater reliability. Agreement refers to the extent to which raters provide essentially the same ratings. It is the degree of interchangeability among raters (Kozlowski & Hattrup, 1992).

Reliability references the proportional consistency of variance among raters. It indicates that the relationship or ranking between rated targets is the same; however, the ratings assigned may differ from rater to rater. It provides an index of consistency.

Fleenor, Fleenor, and Grossnickle (1996) concluded that due to various sources of measurement error, interrater reliability, when reported alone, does not indicate the quality of the ratings. Their findings demonstrated that measures of interrater reliability and interrater agreement are separate indices and that they should be reported together on each set of rating data to provide a more complete evaluation of rating quality. In addition, for clarity and replication, researchers should indicate the methods used when calculating these indices.

2 Interrater Agreement

2.1 Interrater Agreement on Single-Item Measures

Finn (1970) postulated a measure of the within group reliability (r_{WG}) of ten judges rating a single target using a single five-category scale in the equation,

$$r_{WG} = 1 - \frac{s_x^2}{s_{EU}^2} \qquad (1)$$

His approach was to first calculate the random or error variance present in the observed ratings. This is the ratio of the observed variance (s_x^2) to the expected variance (s_{EU}^2). If the expected variance was zero the ratings would demonstrate a rectangular (uniform) distribution. To calculate the expected variance he chose the variance formula for a discrete uniform distribution, $s_{\mathrm{EU}}^2 = (A^2 - 1) / 12$, where A is the number of response categories (Selvage, 1976).

Some have questioned Finn's choice of a discrete uniform distribution (Selvage, 1976). It may not be plausible that all judges would be equally likely to use all possible categories of the scale. It is more likely that the scores would be normally distributed around the mean score. Thus, the selection of a uniform distribution will result in an overestimating of expected variance to the extent of the presence of unimodal response bias, resulting in an overestimating of within-group agreement (Bliese, 2000; Brown & Hauenstein, 2005; Kozlowski & Hattrup, 1992).

The last step in Finn's model was to convert the random or error variance to the non-error variance. Subtracting 1.0 from the error variance yields the non-error variance present in the ratings, which Finn referred to as "a reliability coefficient" (1970, p. 72). A rating of $r_{WG} = 1$ indicates perfect agreement or reliability, 0 reveals the presence of a white-noise-style absence of agreement, and a rating of less than zero indicates systematic disagreement (Harvey & Hollander, 2004). Ratings of less than zero should be set equal to zero (James, Demaree, & Wolf, 1984).

2.2 Interrater Agreement on Multi-Item Measures

James (1984) adapted Finn's equation for the interrater reliability of judges rating a single target using a single item variable, to provide a measure of interrater reliability of judges rating a single target using a multiple item variable. He assumed that the judges mean scores for J items ($J = 1, .., J$) "are 'essentially parallel' indicators of the same construct" (p. 88). James proposed the equation,

$$r_{WG(J)} = \frac{J\left(1 - \frac{\bar{s}_x^2}{s_{EU}^2}\right)}{J\left(1 - \frac{\bar{s}_x^2}{s_{EU}^2}\right) + \frac{\bar{s}_x^2}{s_{EU}^2}} \qquad (2)$$

where the observed variance in Equation 1 was replaced with the average variance of the items in the measurement scale and the Spearman-Brown prophecy formula was added. The result ($r_{WG(J)}$) was the within-group interrater reliability for judges' mean scores based on J essentially parallel items.

Lindell, Brandt, and Whitney (1999) challenged the use of r_{WG} as a measure of reliability. They concluded that since it is incorrect for a measure of reliability to return a value that is less than zero, r_{WG} is a measure of agreement rather than reliability. As a measure of agreement the application of the Spearman-Brown prophecy formula in Equation 2 is improper. They suggested

$$r^{*}_{WG(J)} = 1 - \frac{\overline{s_x^2}}{s_{EU}^2} \qquad (3)$$

as a formula for interrater agreement. They demonstrated that $r^*_{WG(J)}$ is a well behaved index that returned a uniform set of values within the interval of -1.0 to 1.0, especially with a 5-point Likert scale and a sample size of 10 or more.

Cohen, Doveh, and Eick (2001) contended that the argument against the Spearman-Brown correction was not justified. They suggested the continued use of $r_{WG(J)}$ as a measure of homogeneity especially when the number of individuals within each group exceeds 10. Harvey and Hollander (2004) argued that random raters would more likely return a set of values that appeared more normally distributed than uniform. They chose not to use $r^*_{WG(J)}$ in their benchmarking research due to their assumption of an underlying construct in $r_{WG(J)}$ and its substantial history of use in research. Due in part to the fact that $r_{WG(J)}$ and $r^*_{WG(J)}$ return sets of values that lie on different metrics, there is currently no benchmark for interpreting $r^*_{WG(J)}$. Whereas, computed $r_{WG(J)}$ values have been benchmarked with a value of .86 suggesting high but not perfect agreement (James, Demaree, & Wolf, 1993), with values in the .70’s to mid .80’s as sufficient for within-group aggregation (Zohar, 2000), with a value of .74 as a relatively high level of agreement sufficiently justifying aggregation (Judge & Bono, 2000), with a value of .87 as acceptable for aggregation (Dirks, 2000), and with values less than the low-to-mid .90’s indicating a questionable level of agreement (Harvey & Hollander, 2004).

Kozlowski and Hattrup (1992) provided clarity regarding the confusion surrounding the use of r_{WG} and $r_{WG(J)}$ as a measure of reliability. Early writers often used the terms interrater reliability and interrater agreement interchangeably. They noted that it

> is unfortunate that the Finn (1970) formulation and the subsequent work by James et al. (1984) labeled $r_{WG(J)}$ an index of within-group interrater reliability, as this seems to have been a source of some confusion in the literature. In spite of this misleading label, $r_{WG(J)}$ was derived and conceptually defined as an agreement index. (p. 161)

James, Demaree, and Wolf (1993, p. 306) responded that their

> technique was cast as a heuristic form of interrater reliability because [their] derivations build on earlier work by Finn (1970), who had characterized his approach as estimating ‘the proportion of non-error variance in the ratings, a *reliability* [italics added] coefficient’. (p. 72)

Their conclusion to recast $r_{WG(J)}$ as an interrater agreement index preserving the definitions and assumptions pertaining to James et al. (1984) regarding scaling, error distributions, prior illustrations, caveats, and recommendations was in agreement with that of Kozlowski and Hattrup (1992). Cohen, Doveh, and Eick (2001) acknowledged the acceptance and continued use of $r_{WG(J)}$ as a measure of agreement.

3. Interrater Reliability

Bartko (1966) demonstrated that for one-way random effects the intraclass correlation coefficient (ICC), defined as the ratio of variances, can be interpreted as a correlation reliability coefficient. To demonstrate that the ratio of rater variance to total variance is interpretable as a correlation coefficient, he put forth the equation

$$ICC(1) = \frac{MSB - MSW}{MSB + [(C-1)MSW]} \qquad (4)$$

where MSB is the between-group mean square, MSW is the within-group mean square, and C is the number of raters (Bartko, 1976). ICC(1) is an appropriate measure when “each target is rated by a different set of *k* judges, randomly selected from a larger population of judges” (Shrout & Fleiss, 1979, p. 421). This is similar to a group of employees who are evaluated by different supervisors selected from the population of supervisors.

A reliability of 1 indicates perfect agreement, a reliability of 0 indicates no agreement, and reliability values less than zero should be interpreted as zero (Bartko, 1976). Therefore, "the 1 – ICC for intraclass correlation ≥ 0 is interpreted as the percentage of variance due to the disagreement among the raters" (p. 763). Bartko went on to note that a high intraclass correlation reliability coefficient indicates a small within-subjects variance and as such can be interpreted as an indicator of agreement. Alternately, James (1982) observed that it is unknown if a low coefficient "is due to reliable differences among raters (perceivers), interactions between the target of perception and raters, random error, or some combination of these factors" (p. 221). In either case, James put forth perceptual agreement as the chief concern when considering data aggregation. He contended that the intraclass correlation coefficient, as a measure of perceptual agreement, should be used as the primary basis for deciding whether to aggregate data.

In general, ICC(1) reliability results have been benchmarked with values greater than or equal to .80 indicating good agreement (Lih-Jiun et al., 2006), with .93 indicating very strong reliability and .73 as evidence of reliability (Blake et al., 2005), and with values greater than .70 indicating test reliability (Smolla, Valla, Bergeron, Berthiaume, & St-Georges, 2004). Shrout and Fleiss (1979) postulated a minimum acceptable value of .75 or .80 for the reliability coefficient. According to Cicchetti (1994) the reliability of clinical instruments is poor for ICCs below .40, fair for ICCs between .40 and .59, good for ICCs between .60 and .74, and excellent for ICCs between .75 and 1.0. However, Klein et al. (2000) observed that "researchers who use eta-squared or ICC(1) to justify aggregation typically assert that aggregation is warranted if the F-test is statistically significant" (p. 517).

Shrout and Fleiss (1979) observed that when using the ICC researchers are often unaware of the different forms of the coefficient, their different uses, and that researchers often fail to indicate the form that was used. Most commonly ICC(1) is misused in place of ICC(2) or ICC(3) resulting in an underestimation of the true correlation. To estimate ICC(2) Bartko (1966) provided the equation

$$ICC(2) = \frac{MSB - MSW}{MSB + (C-1)MSW + C(MSR - MSE)/R} \qquad (5)$$

where MSE is the mean square of the residual error and R is the number of targets rated. ICC(2) is used when "a random sample of *k* judges is selected from a larger population, and each judge rates each target" (Shrout & Fleiss, 1979, p. 421). This is similar to a group of supervisors selected from the population of supervisors who will each rate each target.

Equation 5 reflects rater agreement rather than consistency. Bartko (1976) argued that consistency was not the correct reliability concept for raters. He held that rater reliability should be limited to agreement. SPSS treats this version of ICC(2) as a two-way random intraclass correlation coefficient having the type value of absolute agreement. ICC(2) and ICC(3) are benchmarked "the same as for their corresponding single-rater reliability index" (Shrout & Fleiss, 1979, p. 426).

Shrout and Fleiss (1979) proposed a third form of intraclass correlation coefficient with the formula

$$ICC(3) = \frac{MSB - MSE}{MSB + [(C-1)MSE]} \qquad (6)$$

ICC(3) is applicable when "each target is rated by each of the same *k* judges, who are the only judges of interest" (p. 421). This is similar to each employee being rated by all of the supervisors, or at least all of the supervisors who ever rate employees. Shrout and Fleiss, and others (see also Algina, 1978) contended that in studies where a single judge or a single group of judges do all of the ratings, the consistency of the ratings should be measured with the judges considered as fixed rather than random effects. SPSS treats this version of ICC(3) as a two-way mixed intraclass correlation coefficient having the type value of consistency.

The most significant limitation related to the use of intraclass correlation coefficients has to do with values on or near 0 (Conway & Schaller, 1998; Glick, 1985; James et al., 1984). Theoretically, scores approaching 0 indicate less consensus. However, scores close to 0 may occur as a result of restricted variability, with a score of exactly 0 indicating perfect consensus. Values closer to 1 are interpreted to indicate greater consensus.

Yet, values on or near 1 can occur as a result of systematic mean differences even when there is no absolute agreement. Finally, intraclass correlation coefficients tend to be descriptively conservative. In many cases intraclass correlation results in a value that fails to adequately convey the magnitude of agreement. As this relates to the various types of intraclass correlation, ICC(1) will usually return a lower measure of reliability than will ICC(2), with ICC(2) returning a lesser degree of reliability than ICC(3).

4. Conclusion

Determining rater agreement and reliability is an essential step in the quantitative study of organizational culture and climate. Without first demonstrating perceptual consistency among raters any conclusions about the nature of an organization’s culture or climate must be viewed with skepticism. Even when measures of rater agreement and reliability are used, a great deal of room exists for alternate interpretations of the data. Organizational culture and climate researchers must ensure that the correct form of the selected index is used, the methods used to calculate the indices are reported, and that justification is provided for the selected benchmarks.

References

Algina, J. (1978). Comment on Bartko's "on various intraclass correlation reliability coefficients". *Psychological Bulletin, 85*(1), 135-138.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 3-11.

Bartko, J. J. (1976). On various Intraclass correlation reliability coefficients. *Psychological Bulletin, 83*(5), 762-765.

Blake, K., Vincent, N., Wakefield, S., Murphy, J., Mann, K., & Kutcher, M. (2005). A structured communication adolescent guide (SCAG): Assessment of reliability and validity. *Medical Education, 39*(5), 482-491.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K.J.Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). Jossey-Bass.

Brown, R. D. & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the rwg indices. *Organizational Research Methods, 8*(2), 165-184.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the rwgj index of agreement. *Psychological Methods, 6*(3), 297-310.

Conway, L. G. & Schaller, M. (1998). Methods for the measurement of consensual beliefs within groups. *Group Dynamics: Theory, Research, and Practice, 2*(4), 241-252.

Dirks, K. (2000). Trust in leadership and team performance: Evidence from NCAA basketball. *Journal of Applied Psychology, 85*(6), 1004-1012.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 71-76.

Fleenor, J. W., Fleenor, J. B., & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. *Journal of Business and Psychology, 10*(3), 367-380.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10*(3), 601-616.

Harvey, R. J. & Hollander, E. (2004). *Benchmarking rwg interrater agreement indices: Let's drop the .70 rule-of-thumb.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*(2), 219-229.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*(2), 306-309.

Judge, T. A. & Bono, J. E. (2000). Five-factor model of personality and transformational leadership. *Journal of Applied Psychology*, *85*(5), 751-765.

Klein, K. J., Bliese, P. D., Kozolowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A. et al. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K.J.Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 512-553). Jossey-Bass.

Kozlowski, S. W. J. & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*(2), 161-167.

Lih-Jiun, L., Ching-Lin, H., Sing-Kai, L., Su, L., Mao-Hsiung, H., & Jau-Hong, L. (2006). Psychometric properties of the modified Emory Functional Ambulation Profile in stroke patients. *Clinical Rehabilitation, 20*(5), 429-437.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23*(2), 127-135.

Selvage, R. (1976). Comments on the analysis of variance strategy for the computation of intraclass reliability. *Educational and Psychological Measurement, 36*, 605-609.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Smolla, N., Valla, J. P., Bergeron, L., Berthiaume, C., & St-Georges, M. (2004). Development and reliability of a pictorial mental disorders screen for young adolescents. *Canadian Journal of Psychiatry, 49*(12), 828-837.

Zohar, D. (2000). A group-level model of safety climate: testing the effect of group climate on microaccidents in manufacturing jobs. *Journal of Applied Psychology, 85*(4), 587-596.