frontiers in Psychology

REVIEW
published: 12 May 2017
doi: 10.3389/fpsyg.2017.00777

# An Overview of Interrater Agreement on Likert Scales for Researchers and Practitioners

*Thomas A. O'Neill* *

*Individual and Team Performance Lab, Department of Psychology, University of Calgary, Calgary, AB, Canada*

Applications of interrater agreement (IRA) statistics for Likert scales are plentiful in research and practice. IRA may be implicated in job analysis, performance appraisal, panel interviews, and any other approach to gathering systematic observations. Any rating system involving subject-matter experts can also benefit from IRA as a measure of consensus. Further, IRA is fundamental to aggregation in multilevel research, which is becoming increasingly common in order to address nesting. Although, several technical descriptions of a few specific IRA statistics exist, this paper aims to provide a tractable orientation to common IRA indices to support application. The introductory overview is written with the intent of facilitating contrasts among IRA statistics by critically reviewing equations, interpretations, strengths, and weaknesses. Statistics considered include $r_{wg}$, $r^*_{wg}$, $r'_{wg}$, $r_{wg(p)}$, average deviation (*AD*), $a_{wg}$, standard deviation ($S_{wg}$), and the coefficient of variation ($CV_{wg}$). Equations support quick calculation and contrasting of different agreement indices. The article also includes a "quick reference" table and three figures in order to help readers identify how IRA statistics differ and how interpretations of IRA will depend strongly on the statistic employed. A brief consideration of recommended practices involving statistical and practical cutoff standards is presented, and conclusions are offered in light of the current literature.

**Keywords: interrater agreement, rwg, multilevel methods, data aggregation, within-group agreement, reliability**

OPEN ACCESS

***Edited by:***
*Con Stough,*
*Swinburne University of Technology, Australia*

***Reviewed by:***
*M. Teresa Anguera,*
*University of Barcelona, Spain*
*Joanne H. Gavin,*
*Marist College, USA*

***Correspondence:**
*Thomas A. O'Neill*
*toneill7@gmail.com*

***Specialty section:***
*This article was submitted to Organizational Psychology, a section of the journal Frontiers in Psychology*

***Received:*** *04 March 2017*
***Accepted:*** *26 April 2017*
***Published:*** *12 May 2017*

***Citation:***
*O'Neill TA (2017) An Overview of Interrater Agreement on Likert Scales for Researchers and Practitioners. Front. Psychol. 8:777. doi: 10.3389/fpsyg.2017.00777*

## INTRODUCTION

The assessment of interrater agreement (IRA) for Likert-type response scales has fundamental implications for a wide range of research and practice. One application of IRA is to quantify consensus in ratings of a target, which is often crucial in job analysis, performance assessment, employment interviews, assessment centers, and so forth (e.g., Brutus et al., 1998; Lindell and Brandt, 1999; Walker and Smither, 1999; Morgeson and Campion, 2000; Harvey and Hollander, 2004). Another application of IRA is to determine the appropriateness of averaging individual survey responses to the group level (van Mierlo et al., 2009). In that spirit, IRA has been used to support the aggregation of individual ratings to the team level, follower ratings of leadership to the leader level, organizational culture ratings to the organizational level, and leadership ratings to the leader level (see discussions by Rousseau, 1985; Chan, 1998; Kozlowski and Klein, 2000). If consensus in the ratings of a target is low, then the mean rating may be a misleading or inappropriate summary of the underlying ratings (George, 1990; George and James, 1993). Underscoring the importance of IRA statistics is that, unlike interrater reliability and consistency statistics, IRA provides a single value of agreement for each rating target, thereby facilitating identification of units of raters who are very high or very low in agreement. This advantageous feature also permits subsequent investigation of other substantive and theoretically interesting

variables that may be related to variance in agreement (Klein et al., 2001; Meade and Eby, 2007), or as a moderator of predictor-criterion relations (e.g., climate strength; Schneider et al., 2002).

IRA are particularly common when collecting systematic observations of behavior or phenomena. For example, Bernardin and Walter (1977) found that training and diary keeping reduced the errors in performance ratings. O'Neill and Allen (2014) investigated subject-matter experts' ratings of product innovation. Weingart et al. (2004) observed and coded negotiation behavior between teams and reported on methods for doing so. Many more examples exist. The key is that IRA becomes highly relevant when judges observe and provide ratings of behavior or phenomena, and the absolute agreement of those ratings is of interest.

Despite the widespread application of IRA statistics and the extensive research focusing on IRA, it appears that considerable challenges persist. For example, a recent review by Biemann et al. (2012) identified situations in which applications of IRA for aggregation of leadership ratings has been misused, as ratings were aggregated (or not) based on flawed interpretations of IRA. A possible contributing factor of the potential for IRA misuse is that considerations of the logic underlying equations and interpretations of alternative IRA statistics have been relatively scattered across organizational (e.g., Lindell and Brandt, 1999), methodological (e.g., Cohen et al., 2001), and measurement (e.g., Lindell, 2001) journals, thereby making it difficult for researchers and practitioners to contrast the variety of statistics available and to readily apply them appropriately. LeBreton and Senter (2008) provided a seminal review of IRA and consistency statistics, but the focus was largely on implications of these types of statistics for multilevel research methods and not on the many other applications of IRA (e.g., agreement in importance ratings collected in job analysis; Harvey, 1991). Elsewhere, IRA statistics have been investigated as dispersion measures of substantive constructs in multilevel research in terms of criterion validity (Meade and Eby, 2007), power (e.g., Roberson et al., 2007), significance testing (e.g., Cohen et al., 2009; Pasisz and Hurtz, 2009), and performance under missing data conditions (Allen et al., 2007; Newman and Sin, 2009). Importantly, some existing articles may be seen as highly technical for some scholars that are new to the IRA literature (e.g., Lindell and Brandt, 1999; Cohen et al., 2001), and other reviews tend to focus on only one or two IRA statistics (e.g., Castro, 2002).

Given the above, what is needed is a relatively non-technical and tractable orientation to IRA that facilitates comparison and interpretation of various statistics for scholars. Accordingly, the purpose of this article is to contribute by providing an accessible and digestible IRA resource for researchers and practitioners with a diverse range of training and educational backgrounds who need to interpret or report on IRA. The current article fills a gap by reporting on an introductory comparative analysis involving eight IRA statistics: $r_{wg}$, $r^*_{wg}$, $r'_{wg}$, $r_{wg(p)}$, average deviation ($AD$), $a_{wg}$, standard deviation ($S_{wg}$), and the coefficient of variation ($CV_{wg}$). A unique contribution is a "quick reference" table containing citations, formulas, interpretations, strengths, and limitations (see **Table 1**). The aim of **Table 1** is to support expedient consideration of the appropriateness of various IRA statistics given a researcher or practitioner's unique situation, and to serve as a foundation for more focused, complex issues addressed in technical guides (e.g., Burke and Dunlap, 2002). Further, three figures attempt to clarify the behavior of IRA statistics and to supplement understanding and interpretation of various IRA statistics. The article introduces James et al.'s (1984) $r_{wg}$, some potential issues with interpretations of that statistic, and numerous contemporary alternatives. Before beginning, a comment on IRA and interrater consistency is offered.

## JAMES ET AL.'S IRA: $r_{wg}$ FOR SINGLE AND MULTIPLE ITEMS

### General Logic

For use on single-item scales, James et al. (1984; see also Finn, 1970) introduced the commonly-used, and perhaps most ubiquitous, IRA statistic known as $r_{wg}$. This statistic is a function of two values: the observed variance in judges' ratings (denoted as $S^2_x$), and the variance in judges' ratings if their ratings were random (denoted as $\sigma^2_e$ in its general form, referred to as the *null distribution*). What constitutes a reasonable standard for random ratings is highly debated. One option, apparently the default in most research, is the rectangular or uniform distribution calculated with the following (Mood et al., 1974):

$$\sigma^2_{eu} = (A^2 - 1)/12 \quad (1)$$

where $A$ is the number of discrete Likert response alternatives. This distribution yields the variance obtained if each Likert category had an equal probability of being selected. Observed variance in judges' ratings on a single item can be compared to this index of completely random responding to determine the proportion of error variance present in the ratings:

$$\text{proportion of random variance in judges' ratings} = S^2_x/\sigma^2_{eu} \quad (2)$$

If this value—the proportion of error variance in judges' ratings—is subtracted from 1, the remaining variance can be interpreted as the proportion of variance due to agreement. Hence, the IRA for single item scales can be:

$$r_{wg} = 1 - (S^2_x/\sigma^2_{eu}) \quad (3)$$

Whereas, Equation (3) is for single-item scales, James et al. (1984) derived an index for multi-item response scales denoted as $r_{wg(j)}$. It applies the Spearman-Brown prophecy formula (see Nunnally, 1978) to estimate IRA given a certain number of scale items (although James et al., 1984 did not use the Spearman-Brown in its derivation; see also LeBreton et al., 2005). Further, the term $S^2_x$ from Equation (3) is substituted with the mean $S^2_x$ derived from judges' ratings on each scale item to yield the following:

$$r_{wg(j)} = J(1 - \overline{S^2_x}/\sigma^2_{eu})/[J(1 - \overline{S^2_x}/\sigma^2_{eu}) + (\overline{S^2_x}/\sigma^2_{eu})] \quad (4)$$

where $\sigma^2_{eu}$ is the same as in Equation (1), and $J$ is the number of items.

**TABLE 1 | Summary of interrater agreement statistics for likert-type response scales.**

| Statistic (citations) | Formula | Interpretation | Strengths | Limitations |
|---|---|---|---|---|
| $r_{wg}$ (James et al., 1984; see also Finn, 1970) | $1 - (S_x^2/\sigma_{eu}^2)$<br>$S_x^2$ = observed variance in judges' ratings on the single item; and $\sigma_{eu}^2$ = variance of the rectangular, uniform null distribution, $(A^2 - 1)/12$, where $A$ is the number of discrete Likert-type response options. | • A value of 1.0 indicates complete agreement.<br>• A value of 0 indicates agreement equal to the null distribution (i.e., one index of completely random responding.<br>• Values below 0 or above 1.0 are assumed to be the result of sampling error and should be reset to 0 (see James et al., 1984). | • Commonly used in the literature and generally known to researchers and reviewers.<br>• Likely the most researched agreement statistic.<br>• Linear function facilitates interpretation. | • Uniform distribution may inappropriately model random responding, and selecting an alternative null distribution can be difficult (for guidance, see LeBreton and Senter, 2008).<br>• May not be directly comparable (i.e., equivalent) across different means of group ratings, number of raters, or sample sizes.<br>• It is not uncommon for values to exceed +1.0 or fall below 0. These inadmissible values might not be the result of sampling error. Resetting the values to 0 may therefore be inappropriate and result in loss of information (Brown and Hauenstein, 2005). |
| $r_{wg(j)}$ (James et al., 1984) | $\frac{J(1-\overline{S_x^2}/\sigma_{eu}^2)}{J(1-\overline{S_x^2}/\sigma_{eu}^2)+(S_x^2/\sigma_{eu}^2)}$<br>$J$ = number of items;<br>$S_x^2$ = mean of the observed variance in judges' ratings on each scale item; and $\sigma_{eu}^2$ = see above. | • A value of 1.0 indicates complete agreement.<br>• A value of 0 indicates agreement equal to the null distribution.<br>• Values below 0 or above 1.0 are assumed to be the result of sampling error and should be reset to 0 (see James et al., 1984). | • Commonly used in the literature and generally known to researchers and reviewers.<br>• Likely the most researched agreement statistic. | • Same as $r_{wg}$, above.<br>• May not be directly comparable (i.e., equivalent) across different means of group ratings or the number of raters.<br>• It is upwardly influenced by the number of discrete Likert scale response options.<br>• Values in between 1.0 and 0 are difficult to interpret because the function is non-linear. |
| $r^*_{wg}$ (Lindell and Brandt, 1997) | $1 - (S_x^2/\sigma_{eu}^2)$<br>or<br>$1 - (S_x^2/\sigma_{mv}^2)$<br>$S_x^2$ = see above; $\sigma_{eu}^2$ = see above; and $\sigma_{mv}^2$ = variance of the maximum dissensus distribution, $0.5(X^2_U + X^2_L) - [0.5(X_U + X_L)]^2$ | • If using $\sigma_{eu}^2$, the interpretation is the same as $r_{wg}$, described above.<br>• If using $\sigma_{mv}^2$, a value of 1.0 indicates complete agreement;.5 indicates agreement equal to the uniform null distribution; and 0 indicates theoretical maximum dissensus.<br>• $r^*_{wg}$ using $\sigma_{mv}^2$ will tend to be greater than is $r^*_{wg}$ using $\sigma_{eu}^2$ and $r_{wg}$ will always be less than is $r^*_{wg}$.<br>• Values below 0 (using $\sigma_{eu}^2$) and below 0.5 (using $\sigma_{mv}^2$) are possible when agreement is low (i.e., it suggests bimodal distributions). | • Presents a compelling alternative to the uniform null distribution ($\sigma_{eu}^2$) by positing the theoretical maximum dissensus ($\sigma_{mv}^2$) for use as a random error term.<br>• Circumvents problems of inadmissible values by allowing for meaningful interpretations when $S_x^2$ it exceeds $\sigma_{eu}^2$. | • May not be directly comparable (i.e., equivalent) across different means of group ratings.<br>• Maximum dissensus may inappropriately model random responding, and selecting an alternative null distribution can be difficult (for guidance, see LeBreton and Senter, 2008).<br>• May be positively correlated with group mean extremity. |

*(Continued)*

**TABLE 1 | Continued**

| Statistic (citations) | Formula | Interpretation | Strengths | Limitations |
|---|---|---|---|---|
| $r^*_{wg(j)}$ (Lindell et al., 1999) | $r^*_{wg(j)} = 1 - (\bar{S}_x^2 / \sigma_{eu}^2)$ or $r^*_{wg(j)} = 1 - (S_x^2 / \sigma_{mv}^2)$ $S_x^2$ = see above; $\sigma^2_{\bar{e}u}$ = see above; and $\sigma_{mv}^2$ = see above. | • Same as $r^*_{wg}$, above. | • Same as $r^*_{wg}$, above.<br>• With increasing items the function remains linear, unlike $r_{wg(j)}$. | • Same as $r^*_{wg}$, above. |
| $r'_{wg(j)}$ (Lindell, 2001) | $1 - (S_y^2 / \sigma_{eu}^2)$ $S_y^2$ = variance of individual judges' scale means; and $\sigma_{eu}^2$ = see above. | • Less attenuated than is $r^*_{wg(j)}$ with $\sigma_{eu}^2$ relative to $r_{wg(j)}$.<br>• Interpretation is otherwise similar to $r^*_{wg(j)}$. | • Less attenuated than is $r^*_{wg}$; Otherwise the strengths are the same as those of $r^*_{wg(j)}$. | • Shares many of the same limitations as does $r^*_{wg(j)}$ except $r'_{wg(j)}$ will often be less attenuated.<br>• Application has been rare in the literature and, accordingly, researchers and reviewers may be unaware of the underlying logic. |
| $r_{wg(p)}$ (LeBreton et al., 2005; LeBreton and Senter, 2008) | • Identify subgroups, calculate each subgroup's agreement score, check homogeneity of variances and, if supported, substitute sample-weighted average group variance (denoted $S^{2x.\tau}$) value into $r_{wg}$ or $r_{wg(j)}$ equation.<br>• Homogeneity of variances can be tested using Fisher's *F*-test by dividing the larger subgroup variance by the smaller subgroup variance, which is approximately distributed as the F distribution with degrees of freedom for subgroup 1/degrees of freedom for subgroup 2 (see Crawley, 2007, p. 289 for application in *R*). | • Has same interpretation as does previous $r_{wg}$ conceptualization except considers subgroup agreement differences by averaging them. | • Allows for consideration of theoretically meaningful subgroups.<br>• Addresses limitation of inadmissible values that can be problematic for $r_{wg}$ and $r_{wg(j)}$. | • Has many of the same interpretational problems as do previous $r_{wg}$ statistics reviewed (e.g., difficulties in choosing an appropriate null distribution).<br>• Can be difficult to generate theoretical predictions *a priori* about the existence of subgroups.<br>• Assumes homogeneity of subgroup variances. If homogeneity assumptions cannot be supported, separate $r_{wg}$ values based on subgroups could be another option. |
| $AD_{M(j)}$ (Burke et al., 1999; Burke and Dunlap, 2002) | $\sum(\lvert x_i - \bar{x}\rvert)/k$ $x_i$ = a judge's rating on the item; $\bar{x}$= is the group mean rating on the item; and *k* is the number of judges. | • Indexes the average distance of judges' ratings from the group's scale mean.<br>• Considerable justification for practical cutoff criteria have been proposed, but they are not without assumptions (see Section Standards for Agreement). | • Interpretation is not complicated by changes (e.g., non-linearity) in the number of Likert categories (bearing in mind greater deviations are expected given category increases).<br>• Circumvents problems associated with choosing an appropriate null distribution. | • May be negatively correlated with group mean extremity.<br>• Does not permit explicit modeling of random responding (i.e., has no null distribution term).<br>• AD values are highly dependent on the number of scale categories employed. This makes it very difficult to compare AD values of scales differing in length. |
| $AD_{M(J)}$ (Burke et al., 1999; Burke and Dunlap, 2002) | $\sum AD_{M(j)}/J$ $J$ = see above. | • Shares interpretations of $AD_{M(j)}$ except generalizes to multi-item scales. | • Same advantages as $AD_{M(j)}$.<br>• Takes the average of each $AD_{M(j)}$ and, therefore, does not unnecessarily complicate the multi-item interpretation. | • Same limitations of $AD_{M(j)}$. |

*(Continued)*

**TABLE 1 | Continued**

| Statistic (citations) | Formula | Interpretation | Strengths | Limitations |
|---|---|---|---|---|
| $a_{wg(1)}$ (Brown and Hauenstein, 2005) | $1 - [(2 * S_x^2)/S_{mpv/m}^2]$<br>$S_x^2$ = see above, and<br>$S_{mpv/m}^2 = [(H+L)M-(M^2) - H*L)*[k/(k-1)]$<br>where H = maximum discrete scale value;<br>L = minimum discrete scale value;<br>$M$ = observed mean rating; and<br>$k$ = number of raters. | • A value of +1.0 indicates perfect agreement, given the group mean.<br>• A value of 0 indicates the observed variance is 50% of the maximum variance, given the group mean.<br>• A value of −1.0 indicates maximum disagreement given the group mean. Will equal single-item $r_{wg}$ when the group mean is at the scale mid-point and the variance equations (sample vs. population) are not mismatched for $r_{wg}$.<br>• Will equal single and multi-item $r^*_{wg}$ using $\sigma_{eu}^2$ when the group mean is at the midpoint and the variances are not mismatched. | • Controls for the extremeness of the group mean by not relying on a single specification of the null distribution.<br>• Uses the unbiased, sample variance to calculate observed and theoretical random variance terms, whereas the $r_{wg}$ family of statistics confound these.<br>• Circumvents problems of inadmissible values.<br>• Will not be affected by sample size because it employs matched variances. | • Requires at least $A$-1 raters for calculating interpretable $a_{wg}$ values, where $A$ is equal to the number of Likert response categories (see Brown and Hauenstein, 2005).<br>• Is not interpretable at face value beyond certain extreme group means. That is, the minimum mean with interpretable $a_{wg} = [L(k-1) + H]/k$; and the maximum mean with interpretable $a_{wg} = [H(k-1) + L]/k$. |
| $a_{wg(j)}$ (Brown and Hauenstein, 2005) | $\sum a_{wg(1)}/J$<br>$J$ = see above. | • Shares interpretations of $a_{wg(1)}$, except generalizes to multi-item scales. | • Same advantages as $a_{wg(1)}$.<br>• Takes the average of each $a_{wg(1)}$ and, therefore, does not unnecessarily complicate the multi-item interpretation. | • Same limitations as $a_{wg(1)}$. |
| $S_{wg}$ (Schmidt and Hunter, 1989) | $\{[\sum (x_i - \bar{x})^2]/(n-1)\}^{1/2}$<br>$x_i$ = a judge's rating on the item, $\bar{x}$ = the group mean rating on the item; and $n$ is the number of group members. | • The root of the average squared judge deviation from the mean. | • Provides a straightforward and direct index of agreement. | • Will be scale dependent such that a greater number of response options will tend to produce greater $S_{wg}$.<br>• Does not permit explicit modeling of random responding (i.e., has no null distribution term). |
| $S_{wg(j)}$ | $\sum S_{wg}/J$<br>$J$ = see above. | • Shares interpretations of $a_{wg(1)}$, except generalizes to multi-item scales. | • Same advantages as $S_{wg}$.<br>• Takes the average of each $S_{wg(1)}$ and, therefore, does not unnecessarily complicate the multi-item interpretation. | • Same limitations as $S_{wg}$. |

*(Continued)*

**TABLE 1 | Continued**

| Statistic (citations) | Formula | Interpretation | Strengths | Limitations |
|---|---|---|---|---|
| $CV_{wg}$ (Allison, 1978; Bedeian and Mossholder, 2000) | $S_{wg}/\bar{x}$<br>s = see above; and $\bar{x}$ = see above. | • Rescales the standard deviation by taking into account the mean.<br>Large values suggest large variance relative to the mean (and scale). | • Samples with larger means may be expected to have greater standard deviations than samples with smaller means. The $CV_{wg}$ will not be affected by the scale mean, thereby facilitating comparisons across samples (i.e., groups) with different means (and scaling). | • It is difficult to decide what constitutes high and low consensus based on $CV_{wg}$ values; therefore, application and interpretation of $CV_{wg}$ may be difficult.<br>• The assumption of a non-negative ratio scale may not always be tenable.<br>• The $CV_{wg}$ is intended for situations in which means vary widely. If groups tend not to differ much on sample means there is little reason to adopt $CV_{wg}$.<br>• Does not permit explicit modeling of random responding (i.e., has no null distribution term). |
| $CV_{wg(j)}$ | $\sum CV_{wg}/J$<br>J = see above. | • Shares interpretations of $CV_{wg}$, except generalizes to multi-item scales. | • Same advantages as $CV_{wg}$. | • Same disadvantages as $CV_{wg}$. |

## Interpretation

**Figure 1** shows the range of $r_{wg}$ values across all possible levels of mean $S_x^2$ based on four raters and a five-point Likert scale (see also Lindell and Brandt, 1997). One observation from **Figure 1** is that the single-item $r_{wg}$ is a linear function, such that complete agreement equals 1.0 and uniform disagreement equals 0 (i.e., raters select response options completely at random). But notice that for $S_x{}^2 > 2$—that is, where $S_x{}^2$ exceeds $\sigma^2_{eu} - r_{wg}$ takes on negative values. **Figure 1** also contains the $r_{wg(j)}$ function ranging from −1.0 to +1.0 across levels of $S_x{}^2$ based on four raters, a five-point Likert scale, and two, five, and ten items. Consistent with expectations, when $r_{wg(j)}$ is 1.0 agreement is perfect and when $r_{wg(j)}$ is 0 there exists uniform disagreement. However, at all other levels of $r_{wg(j)}$, interpretation is complicated because the shape of the function changes depending on the number of items. Consider that, as the mean $S_x{}^2$ moves from 0 to 1.5, $r_{wg(j)}$ ranges from 1.0 to 0.40 [$r_{wg(2)}$], 1.0 to 0.63 [$r_{wg(5)}$], and 1.0 to 0.77 [$r_{wg(5)}$] suggesting that $r_{wg(j)}$ is insensitive to substantial changes at reasonable levels of mean $S_x^2$, and it might imply surprisingly high agreement even when there is considerable variance in judges' ratings. This also illustrates the extent to which the problem increases in severity as the number of items increases. The pattern creates the potential for misleading or inaccurate interpretations when the shape of the function is unknown to the researcher. Another issue is that $S_x{}^2 > 2$ produces inadmissible values that are outside the boundaries of $r_{wg(j)}$ (i.e., $< 0$ or $> 1.0$). Regarding inadmissible values, James et al. (1984) suggested that these may be a result of sampling error. Other possible contributing factors include inappropriate choices of null distributions and the existence of subgroups. One recommended procedure is to set inadmissible values to 0 (James et al., 1993). This could be an undesirable heuristic, however, because it results in lost information (Lindell and Brandt, 1999, 2000; Brown and Hauenstein, 2005).

## Potential Cause for Concern

Whereas, $r_{wg}$ is arguably the most widely used IRA statistic, there are five issues concerning its interpretation. First, there is the issue of non-linearity described above. This non-linearity, occurring with increased magnitude as the number of scale items increases, renders interpretations of agreement levels ambiguous compared to interpretations of linear functions. The appropriateness of interpretations may be particularly weak if the researcher or practitioner is unaware that the function is non-linear. Indeed, scales with a large number of items will almost always have very high agreement (Brown and Hauenstein, 2005; cf. Lindell and Brandt, 1997; Lindell et al., 1999; Lindell, 2001), which limits the interpretational and informative value of $r_{wg(j)}$ with scales containing more than a few items. **Figure 1** clarifies this. Second, there are difficulties involving inadmissible values, also described above. Resetting these values to 0 or 1.0 seems suboptimal because potentially useful information is arbitrarily discarded. It would be advantageous if that additional information could be used to further shed light on agreement. Third, $r_{wg}$ and $r_{wg(j)}$ appear to be related to the mean rating extremity. Brown and Hauenstein (2005) found a correlation

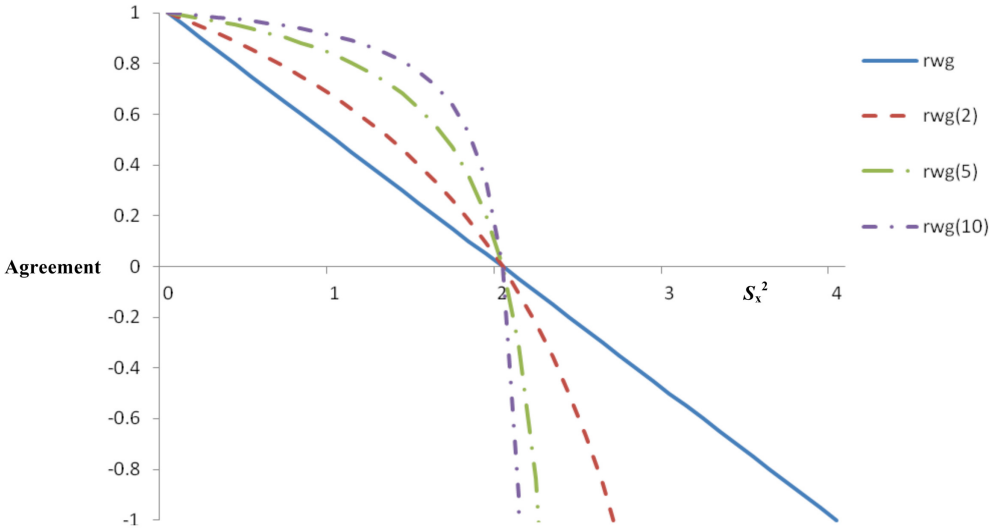


**FIGURE 1 | Single and multiple-item $r_{wg}$ across levels of $S_x^2$ (five-point scale).**

between mean judge ratings and $r_{wg(j)}$ values of 0.63. This is not surprising because mean ratings falling closer to the scale endpoint must have restricted variance (i.e., agreement). Thus, $r_{wg}$ will be affected by the mean rating. Fourth, the typical selection of $\sigma_e^2$, the theoretical distribution of random variance, seems to be the rectangular distribution, described above as $\sigma_{eu}^2$ (Cohen et al., 2009). But the $\sigma_{eu}^2$ uses scaling that leads to inadmissible values (i.e., $r_{wg} < 0$), and other distributions may be an improvement (LeBreton and Senter, 2008). Whereas, James et al. (1984) offered alternatives to $\sigma_{eu}^2$ that attempt to model response tendencies or biases, in many cases it is difficult to make a choice other than $\sigma_{eu}^2$ that can be defended (for laudable attempts, see Kozlowski and Hults, 1987; LeBreton et al., 2003). One alternative to $\sigma_{eu}^2$, suggested by Lindell and Brandt (1997), however, seems promising (described further below). Fifth, the observed variance in the numerator, $S_x^2$, tends to decrease with sample size, which creates the potential to spuriously increase $r_{wg}$ (Brown and Hauenstein, 2005).

Given the above issues involving James et al.'s (1984) $r_{wg}$, the remainder of this article describes some alternatives and how each alternative was proposed to address at least one of the issues raised. Knowledge of this is intended to help the researcher or practitioner make informed decisions regarding the most applicable statistic (even $r_{wg}$) given his or her unique situation.

## $r^*_{wg}$ WITH THE RECTANGULAR NULL AND MAXIMUM DISSENSUS NULL DISTRIBUTIONS

### General Logic

In order to overcome shortcomings of non-linearity and inadmissible values of $r_{wg}$ and $r_{wg(j)}$, Lindell et al. (1999) proposed $r^*_{wg}$. $r^*_{wg}$ using $\sigma_{eu}^2$ is equal to $r_{wg}$ except $r^*_{wg}$ allows for meaningful negative values to −1.0. Negative values will occur when $S_x^2$ exceeds the variance of the rectangular distribution, $\sigma_{eu}^2$, and these negative values indicate bimodal distributions. In other words, clusters of raters are at or near the scale end points. Unlike $r_{wg}$, which does not consider negative values to be admissible, $r^*_{wg}$ recognizes that this information can provide theoretical insight into the nature of the disagreement. $r^*_{wg(j)}$ with $\sigma_{eu}^2$ also uses the same equation as does $r_{wg}$ but instead uses the mean variance in the numerator:

$$r^*_{wg} = 1 - (\overline{S_x^2}/\sigma_{eu}^2) \quad (5)$$

where $S_x^2$ is the mean of the item variances of judge ratings. **Figure 2** illustrates that $r^*_{wg(j)}$ has the favorable property of linearity, meaning that it will not be affected by increasing scale items. Lindell et al. (1999) suggested that interpretation may be aided by keeping the range of admissible values to those of James et al.'s (1984) $r_{wg}$ and $r_{wg(j)}$ (i.e., 0–1.0). Lindell et al. (1999) pointed out that this could be done by setting the expected random variance, $\sigma_e^2$, to the maximum possible disagreement, known as maximum dissensus. Maximum dissensus ($\sigma_{mv}^2$) is:

$$\sigma_{mv}^2 = 0.5(X^2_U + X^2_L) - [0.5(X_U + X_L)]^2 \quad (6)$$

where $X_U$ and $X_L$ are the upper and lower discrete Likert categories, respectively (e.g., "5" and "1" on a five-point scale; Lindell, 2001). Maximum dissensus occurs when all judges are distributed evenly at the scale endpoints, and it can be used in the denominator of the $r^*_{wg}$ or $r^*_{wg(j)}$ equations. For example, for multi-item scales:

$$r^*_{wg(j)} = 1 - (\overline{S_x^2}/\sigma_{mv}^2) \quad (7)$$

It is instructive to point out that on a five-point scale, $\sigma_{eu}^2$ is 2 and $\sigma_{mv}^2$ is 4. Thus, the use of maximum dissensus essentially

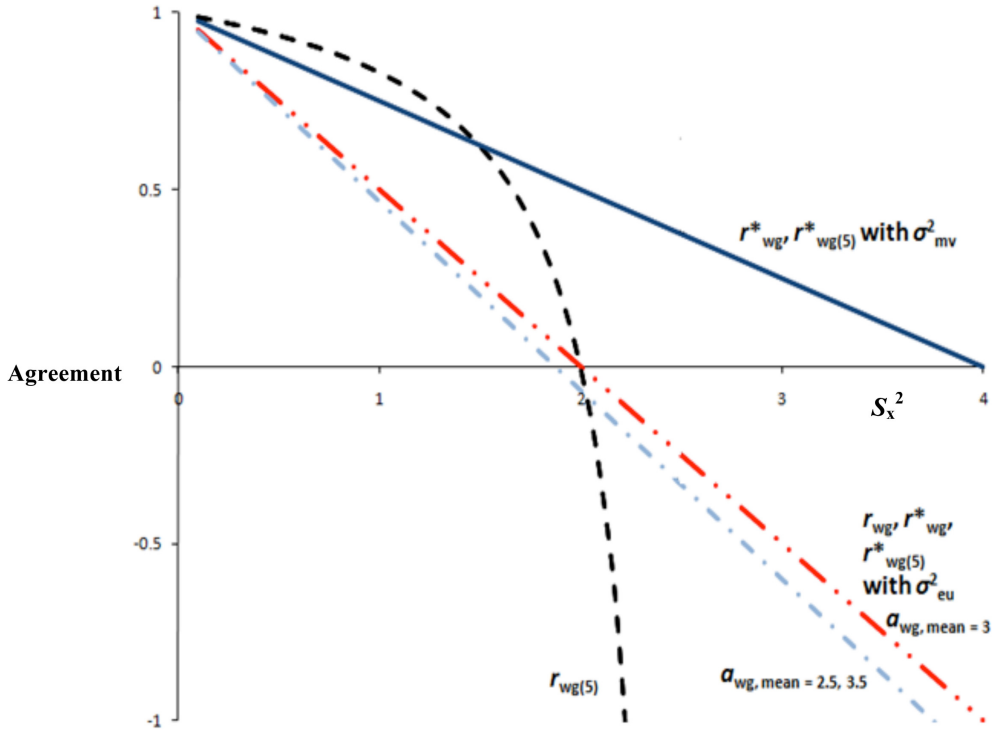


**FIGURE 2 | Sample $r_{wg}$-family and $a_{wg}$ statistics across levels of $S_x^2$ (five-point scale).**

rescales James et al.'s (1984) $r_{wg}$ such that all values of ${S_x}^2$ will result in $r^*_{wg}$ values within the range of 0 and 1.0. This index avoids the problem of non-linearity and corresponding inflation potential of $r_{wg(j)}$ and addresses the problem of inadmissible values.

## Interpretation

**Figure 2** contains functions for $r^*_{wg}$ and $r^*_{wg(j)}$ with ${\sigma_{eu}}^2$ and $\sigma^2_{mv}$. Values for $r^*_{wg}$ and $r^*_{wg(j)}$will range from −1.0 to 1.0 if the denominator is ${\sigma_{eu}}^2$, wherein a value of 0 is uniform disagreement (i.e., ${S_x}^2 = {\sigma_{eu}}^2$) and a value of −1.0 is maximum dissensus (i.e., ${S_x}^2 = {\sigma_{mv}}^2$). Note the advantage of $r^*_{wg}$ and $r^*_{wg(j)}$ in that information is preserved by assigning a meaningful interpretation to negative values. Values for $r^*_{wg}$ and $r^*_{wg(j)}$ will range from 0 to 1.0 when the denominator is ${\sigma_{mv}}^2$, wherein a value of 0.5 is uniform disagreement (i.e., ${S_x}^2 = {\sigma_{eu}}^2$), and a value of 0 is maximum dissensus (i.e., $S^2_x = {\sigma_{mv}}^2$). Taken together, $r^*_{wg}$ and $r^*_{wg(j)}$ potentially address three drawbacks of James et al.'s (1984) statistics. First, negative values are interpretable by incorporating the concept of maximum dissensus. Second, by using $S^2_x$ in the numerator, the multi-item agreement index is not extensively affected by the addition of scale items, which is a major interpretational difficulty of $r^*_{wg(j)}$. Third, $r^*_{wg}$ and $r^*_{wg(j)}$ have the further advantage of avoiding inadmissible values that exceed +1.0.

# FURTHER ADVANCES ON $r^*_{wg}$: DISATTENUATED MULTI-ITEM $r^*_{wg}$: ($r'_{WG(j)}$)

## General Logic

One of the difficulties with Lindell et al.'s (1999) observed $r^*_{wg}$ statistics, described above, is the use of $\sigma^2_{mv}$ when comparisons between James et al.'s (1984) $r_{wg}$ are of interest (Lindell, 2001). The problem lies in the differences in ranges; James et al.'s (1984) $r_{wg}$ statistics have admissible values within 0 and +1.0, whereas for $r^*_{wg}$ statistics that use $\sigma^2_{mv}$ the admissible range is from 1.0 to +1.0. Two steps could be taken to remedy this problem. First, as mentioned above, Lindell et al. (1999) observed that $r^*_{wg}$ statistics could be computed with $\sigma^2_{eu}$. This facilitates comparisons, and also allows the researcher to use a multi-item $r^*_{wg}$ that would have similar behavior compared to single-item $r_{wg}$. But, a further problem noted by Lindell (2001) is that $r^*_{wg}$ and $r^*_{wg(j)}$ with ${\sigma_{eu}}^2$ will be attenuated in comparison to admissible $r_{wg}$ and $r_{wg(j)}$ values. Thus, a second avenue offered by Lindell (2001) to address the relative attenuation of $r^*_{wg(j)}$ using $\sigma^2_{eu}$ is an alternative called $r'_{wg(j)}$. $r'_{wg(j)}$ uses the variance of raters' scale scores on multi-item scales (referred to as ${S_y}^2$, see **Table 1** for derivation details):

$$r'_{wg(j)} = 1 - ({S_y}^2/{\sigma_{eu}}^2) \quad (8)$$

## Interpretation

Lindell (2001) demonstrated that $r'_{wg(j)}$tends to produce larger values than does $r^*_{wg(j)}$ using $\sigma^2_{eu}$, thereby addressing the issue of

attenuation. Otherwise, $r'_{wg(j)}$ has the same general interpretation as does $r^*_{wg(j)}$, although it might be expected to share the limitation of being correlated with group mean extremity. A further difficulty might involve the need to extensively explain $r'_{wg(j)}$ and its logic as reviewers may not be as familiar with this agreement statistic as they are with more frequently employed agreement indices (see **Table 1**).

## POOLED AGREEMENT FOR SUBGROUPS: $r_{wg(p)}$

### General Logic

As a possible remedy for the problem of inadmissible $r_{wg}$ values that fall below 0 or above 1.0, LeBreton et al. (2005) offered $r_{wg(p)}$. The rationale is that inadmissible values suggest bimodal response distributions, and the different clusters comprise subgroups. Therefore, separate IRA could be computed for each subgroup, which could then be pooled. Accordingly, $r_{wg(p)}$ computes the sample-size weighted average of raters' variance for the two groups, and this value is used in James et al.'s $r_{wg}$ or $r_{wg(j)}$ (see also **Table 1**). This will effectively remove the possibility of inadmissible values.

### Interpretation

There are a few noteworthy drawbacks involving the use of $r_{wg(p)}$. Calculating the pooled $r_{wg(p)}$ requires homogeneity of observed variances (e.g., using Fisher's *F*-test; see **Table 1**), otherwise pooling the variances to calculate the $r_{wg(p)}$ may not be justifiable. Another limitation is that these subgroups may be difficult to identify theoretically or *a priori*; thus, capitalization on chance is possible (LeBreton et al., 2005). This can be contrary to purpose as most researchers are interested in a pre-specified set of judges (e.g., team membership). Finally, given that $r_{wg(p)}$ has its basis on $r_{wg}$ and $r_{wg(j)}$, $r_{wg(p)}$ would share many of the limitation of James et al.'s (1984) statistics. Notwithstanding these limitations, $r_{wg(p)}$ does provide a potentially advantageous extension of $r_{wg}$ and $r_{wg(j)}$ for use when subgroups are suspected.

## AVERAGE DEVIATION INDEX

### General Logic

One major difficulty inherent in $r_{wg}$ is the choice of a suitable null distribution. As reviewed above, there is the choice of the rectangular distribution or the maximum dissensus distribution. Moreover, there are other potential distributions, such as skewed bell-shaped distributions, that may more realistically represent null distributions by taking into account factors such as socially-desirable responding or acquiescence tendencies (James et al., 1984; see also discussions by Schmidt and DeShon, 2003; LeBreton and Senter, 2008). Importantly, the selected distribution affects the magnitude of IRA statistics, their interpretation, and comparisons to other IRA statistics. To circumvent difficulties in choosing a null distribution, Burke et al. (1999) offered the average deviation index. The average deviation is calculated by determining the sum of the differences between each rater and the mean rating divided by the number of raters:

$$AD_{M(j)} = \sum(|x_i - \bar{x}|)/k \quad (9)$$

where $AD_{M(j)}$ is the average deviation of judges' ratings on a given item, $x_i$ is a judge's rating on the item, $\bar{x}$ is judges' mean rating on the item, and $k$ is the number of judges. When there are multiple items:

$$AD_{M(J)} = \sum AD_{M(j)}/J \quad (10)$$

where $AD_{M(J)}$ is the average deviation of judges' ratings from the mean judge rating across items, $AD_{M(j)}$ is the average deviation on a given item, and $J$ is the number of scale items. Note that $AD$ can be generalized for use with the median, instead of the mean, in order to minimize the effects of outlier or extreme raters.

### Interpretation

The average deviation approach is advantageous as it provides a direct assessment of IRA without invoking assumptions about the null distribution. Moreover, Burke and Dunlap (2002) made useful inroads for determining cutoffs for supporting aggregation, as they attempt to control for the number of Likert response options by suggesting a cutoff criterion of $A/6$ (where $A$ is the number of Likert categories; cutoff criteria are discussed further below). On the downside, like $r_{wg}$ statistics, the average deviation will be correlated with the group mean such that means closer to the extremities will be negatively related to average deviation values (see **Table 1**). In addition, whereas some forms of $r_{wg}$ can suffer from inadmissible values, AD has the problem of having no standard range whatsoever. Thus, AD values will be difficult to compare across scales with a different numbers of categories.

## BROWN AND HAUENSTEIN'S "ALTERNATIVE" ESTIMATE OF IRA: $a_{wg(1)}$

### General Logic

Brown and Hauenstein (2005) developed the $a_{wg(1)}$ to overcome the limitation of other agreement indices that are correlated with the extremeness of mean ratings. The closer the mean rating is to the scale endpoint (i.e., the extremity of the group mean), the lower the variance in those ratings, and the greater the agreement. This confounds all of the above IRA statistics with the group mean and consequently renders them incomparable across groups with different means. Accordingly, Brown and Hauenstein presented $a_{wg(1)}$, which uses, as a null distribution, the maximum possible variance (i.e., maximum dissensus) given a group's mean:

$$S_{mpv/m}{}^2 = [(H + L)M - (M^2) - H^*L]^*[k/(k-1)] \quad (11)$$

where $S_{mpv/m}{}^2$ is the maximum possible variance given $k$ raters, $M$ is the observed mean rating, and H and L are the maximum

and minimum discrete scale values, respectively. Once the maximum possible variance is known, the single-item $a_{wg}$ is:

$$a_{wg} = 1 - [(2^{*}S_x{}^2)/S_{mpv/m}{}^2] \quad (12)$$

Note that multiplying $S_x{}^2$ by 2 is arbitrary, and is done to give it the same empirical range as James et al.'s (1984) $r_{wg}$. For multi-item scales, the single-item $a_{wg}$s are averaged:

$$a_{wg(j)} = \sum a_{wg(1)}/J \quad (13)$$

## Interpretation

**Figure 2** contains $a_{wg}$ values for means of 3, 2.5, and 3.5, on a five-point scale. Values of −1.0 indicate maximum dissensus (i.e., judge's ratings are on the scale endpoints as much as possible so as to maximize observed variance, $S_x{}^2$), 0 indicate the observed variance is 50% of the maximum variance (i.e., uniform disagreement), and +1.0 indicate perfect agreement, given the group mean. Note that this is the same interpretation as of the single-item $r_{wg}$, except $a_{wg}$ is adjusted for the group mean. Moreover, single and multi-item $a_{wg}$ are linear functions, thereby enhancing ease of interpretation (see **Figure 2**). Notice that $a_{wg}$ for means departing from the midpoint of the scale are slightly lower, thereby taking into account decreases in maximum dissensus as a result of restricted variance. Finally, $a_{wg}$ will not be influenced by sample sizes or number of scale anchors, which are notable additional advantages.

One limitation to Brown and Hauenstein's (2005) $a_{wg}$ is that $S_{mpv/m}{}^2$ cannot be applied when the mean is extreme (e.g., 4.9 on a 5-point scale). This is because $S_{mpv/m}{}^2$ assumes that at least one rater falls on each scale endpoint, although this is impossible given some extreme means. Thus, there are boundaries in means, outside of which appropriate maximum variance estimates should not be applied (Brown and Hauenstein):

$$\text{Minimum mean with interpretable } a_{wg} = [L(k - 1) + H]/k \quad (14)$$

$$\text{Maximum mean with interpretable } a_{wg} = [H(k - 1) + L]/k \quad (15)$$

where L and H are the lowest and highest scale values, and $k$ is the number of judges. This is however, a relatively modest limitation because mean ratings falling beyond these boundaries are likely to indicate strong agreement, as values close to the endpoints will only occur when agreement is high. Nevertheless, $a_{wg}$ scores exceeding interpretational boundaries cannot be compared at face value to other groups' $a_{wg}$s. An additional limitation of $a_{wg}$ is that, unlike most other IRA statistics, $a_{wg}$ is based on more than a single parameter (e.g., the observed variance, $S_x^2$). It also includes the mean. As both $S_x{}^2$ and $\bar{x}$ are affected by sampling error, sampling error may have a greater influence on $a_{wg}$ than on some other IRA statistics (Brown and Hauenstein). Limitations aside, $a_{wg}$ is advantageous because it controls for the mean rating using a mean-adjusted maximum dissensus null distribution and it has a linear function.

Unlike other agreement statistics, $a_{wg}$ matches the variances ($S_x{}^2$, $S_{mpv/m}{}^2$) on whether they employ the unbiased (denominator is $n - 1$) or population-based (denominator is $n$) variance equations. The $r_{wg}$ family mixes unbiased and population-based variances (i.e., $S_x{}^2$, $\sigma_{eu}{}^2$, respectively), thereby potentially leading to inflation of $S_x{}^2$ as sample sizes decreases (see Brown and Hauenstein, 2005). This results in larger values for the $r_{wg}$ family as sample size increases, and, therefore, IRA agreement will almost always be high in large samples (Kozlowski and Hattrup, 1992). Conversely, $a_{wg}$ matches the variances by employing sample-based equations for both of $S_x{}^2$ and $S_{mpv/m}{}^2$, making $a_{wg}$ independent of sample size. If population-level data is obtained, controls for sample size can be employed by substituting $k$ for $k-1$ in both of $S_x{}^2$ and $S_{mpv/m}{}^2$ (Brown and Hauenstein; see **Table 1**).

# STANDARD DEVIATION

## General Logic

The square root of the variance term used throughout the current article, $S_x{}^2$, is the standard deviation, $S_{wg}$. As $S_{wg}$ is the square root of the average squared deviations from the mean, Schmidt and Hunter (1989) advocated for $S_{wg}$ as a straightforward index of IRA around which confidence intervals can be computed. Using the standard deviation addresses problems associated with choosing a null distribution and of non-linearity. The average $S_{wg}$ across items can be used in the case of multi-item scales.

## Interpretation

Advantages of using $S_{wg}$ as an index of IRA is that it is a common measure of variation, and its interpretation is not complicated by the use of multi-item scales or non-linear functions [see $r_{wg(j)}$]. However, the $S_{wg}$ has not always enjoyed widespread application. It cannot be explicitly compared to random response distributions, and this could be of interest. It also tends to increase with the size of the scale response options, meaning that comparisons across scales are not feasible. Finally, it will also tend to decrease with increases in sample size; thus, it will not be sample-size independent.

# COEFFICIENT OF VARIATION

A problem with most IRA statistics reviewed is that they are scale dependent, making comparisons across scales with widely discrepant numbers of Likert response options problematic. With greater numbers of response options, the variance will tend to increase. Thus, the amount of variance (e.g., $S_x{}^2$) could partly depend on scaling, thereby presenting a possible source of contamination for many IRA statistics. One way to address this difficulty is to control for the group mean, because means will typically be larger with greater numbers of response options. One statistic that attempts to address this issue is the coefficient of variation ($CV_{wg}$). The $CV_{wg}$ indexes IRA by transforming the standard deviation into a variance estimate that is less scale dependent, using the following:

$$CV_{wg} = \{[\sum (x_i - \bar{x})^2]/(n - 1)\}^{1/2}/\bar{x} \quad (16)$$

## MULTI-ITEM $CV_{wg}$ COULD BE COMPUTED BY AVERAGING $CV_{wg}$ OVER THE $J$ ITEMS.

### Interpretation

By dividing the standard deviation (the numerator) by the group mean, $CV_{wg}$ aims to provide an index of IRA that is not severely influenced by choice of scale, thereby facilitating comparisons of IRA across different scales. For example, the $CV_{wg}$ for a sample with a standard deviation of 6 and a mean of 100 would be identical to the $CV_{wg}$ for a sample with a standard deviation of 12 and a mean of 200. **Figure 3** contains $CV_{wg}$ for means equal to 50, 100, and 200 with standard deviations ranging from 0 to 15. An inspection of **Figure 3** indicates that the $CV_{wg}$ increases faster with increases in standard deviations for low means than for high means, thereby taking into account the difference in variation that may be related to scaling. Thus, the $CV_{wg}$ could be helpful in comparing IRA across scales with different numbers of response options. On the other hand, it is only helpful for relative (to the mean) comparisons, and not absolute comparisons (Allison, 1978; Klein et al., 2001). This can be clarified by observing that the addition of a constant to a set of scores will affect the mean and not the standard deviation, making it difficult to offer meaningful interpretations of absolute $CV_{wg}$s (but ratio scaling helps; Bedeian and Mossholder, 2000). Another issue is that negative $CV_{wg}$ will occur in the presence of a negative mean, but a negative $CV_{wg}$ is not theoretically interpretable. Thus, a further requirement is non-negative scaling (Roberson et al., 2007).

## STANDARDS FOR AGREEMENT

What constitutes strong agreement within raters? This is an important question as researchers wishing to employ IRA statistics to support and justify decisions. For example, aggregation of individuals' responses to the group (mean) level may assume a certain level of consensus (Chan, 1998). Or, consensus thresholds may be used in the critical incident technique in job analysis, where performance levels of the employees involved in the incident should be agreed upon by experts (Flanagan, 1954). Identifying a unified set of standards for agreement, however, has proven elusive. Two general approaches to identifying standards for agreement have been suggested: statistical and practical. These are considered briefly below.

### Practical Standards

Historically, the emphasis on IRA has been on practical standards or "rules of thumb." For example, the $r_{wg}$ family of statistics has relied on the 0.70 rule of thumb. It is important to acknowledge that the decision to choose 0.70 as the cutoff was based on what amounted to no more than a phone call (see personal communication, February 4th, 1987, in (George, 1990), p. 110; for a discussion, see LeBreton et al., 2003; Lance et al., 2006), and James et al. (1984) likely never intended for this cutoff to be so strongly, and perhaps blindly, adopted. Nevertheless, a perusal of **Figures 1**, **2** clearly shows why common, rule of thumb standards for any of these statistics are difficult to support. A value of 0.70 has a different meaning for most statistics in the $r_{wg}$ family. Further, even within-statistic, different situations may render that statistic incomparable. For example, agreement of 0.70 for an $r_{wg}$ based on 10 items vs. an $r_{wg(j)}$ based on two items is a different agreement benchmark because of the non-linearity. Identification of the null distribution is another influencing factor, as **Figure 2** clearly shows that use of $\sigma_{eu}^2$ vs. $\sigma_{mv}^2$ changes the interpretation of any absolute $r_{wg}$ value (e.g., 0.70), not to mention other potential null distributions (see LeBreton and Senter, 2008).

As noted by Harvey and Hollander (2004), justification for a cutoff of 0.70 is based on an assumption that agreement is similar to reliability, and reliabilities exceeding 0.70 are preferred. However, reliability is about consistency of test scores, not absolute agreement of test scores. Test scores can be perfectly reliable (consistent) but very distinct in absolute quantities. Reliability can, and should be, approached using Generalizability Theory. G-theory involves the systematic investigation of all sources of consistency and error (Cronbach et al., 1972). For example, O'Neill et al. (2015) identified raters as the largest source of variance in performance ratings, rather than rates or dimensions. Thus, drawing on the 0.70 cutoff from reliability theory is not tenable as this rule of thumb underscores the complexity of reliability. An additional assumption is that a single value (e.g., 0.70) would be meaningfully compared across situations and possibly statistics in the $r_{wg}$ family. These assumptions seem untenable, and adopting any standard rule of thumb for agreement involving the entire $r_{wg}$ family would appear to be misguided (see Harvey and Hollander, 2004). For many of the same reasons (i.e., incomparability across different situations), there is no clear avenue for setting practical cutoff criteria for $S_{wg}$ and $CV_{wg}$.

LeBreton and Senter (2008) proposed that standards for interpreting IRA could follow the general logic advanced by Nunnally (1978; see also Nunnally and Bernstein, 1994). Specifically, cutoff criteria should be more stringent when decisions will be highly impactful on the individuals involved (e.g., performance appraisal for administrative decision making). Where applicable, LeBreton and Senter (2008) added that cutoff criteria should consider the nature of the theory underlying aggregation for multilevel research, and the quality of the measure (e.g., newly-established measures may be expected to show lower IRA than do well-established measures). For application to the $r_{wg}$ family, the following standards were recommended: 0–0.30 (lack of agreement), 0.31–0.50 (weak agreement), 0.51–0.70 (moderate agreement), 0.71–0.90 (strong agreement), and 0.91–1.0 (very strong agreement). Whereas, these standards will have different implications and meaning for different types of $r_{wg}$ and null distributions (consider **Figures 1**, **2**), LeBreton and Senter (2008) proposed the standards for all forms of $r_{wg}$. Thus, there is a strong "disincentive" to report versions of $r_{wg}$ that will result in the appearance of lower IRA (e.g., using a normal distribution for the null; LeBreton and Senter, p. 836). Nevertheless, they challenged researchers to select the most appropriate $r_{wg}$ by using theory (especially in the identification of a suitable null distribution), with the hope that professional judgment will prevail. Future research will be telling with regard to whether or not researchers adopt LeBreton and Senter's (2008) recommended practices.

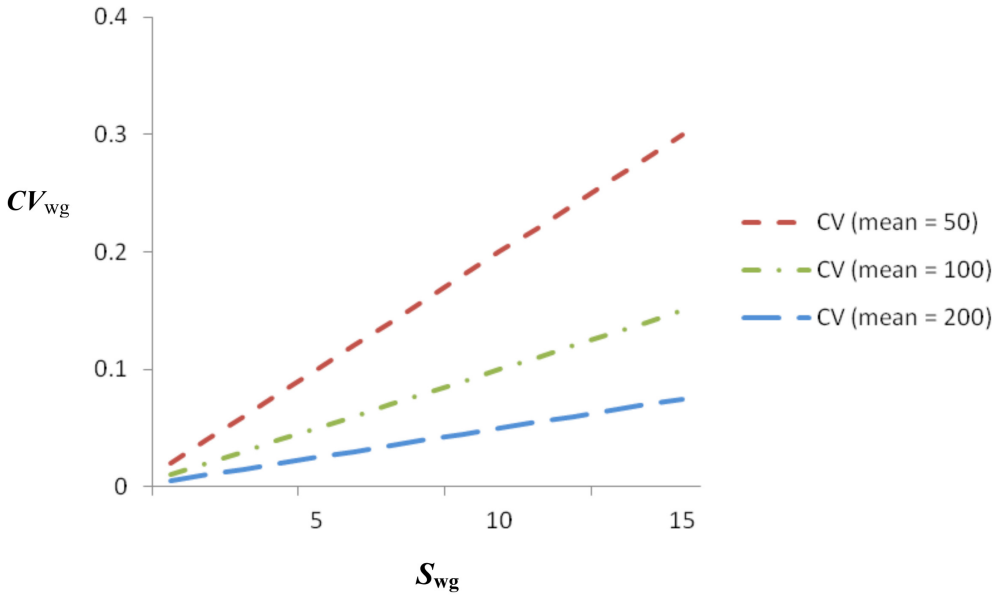


**FIGURE 3 | $CV_{wg}$ from 0 to 15 within-group standard deviations ($S_{wg}$) and means of 50, 100, and 200.**

Turning to other IRA statistics, Burke and Dunlap (2002) suggested that practical significance standards for *AD* could apply the decision rule *A*/6 (where *A* is the number of Likert categories). Thus, for a five point Likert scale 5/6 = 0.83, and *AD* values exceeding 0.83 would be seen as not exhibiting strong agreement. But this decision rule makes two assumptions in its derivation (see Burke and Dunlap for details): (a) the basis is in classical test theory and that interrater reliability should exceed 0.70; and (b) the appropriate null distribution is the rectangular distribution. If these assumptions can be accepted, then the AD has a sound approach for determining cutoffs for practical significance. But if that "null distribution fails to model disagreement properly, then the interpretability of the resultant agreement coefficient is suspect" (Brown and Hauenstein, 2005, p. 166). Elsewhere, Burke et al. (1999) proposed different criteria. They suggested that *AD* should not exceed 1.0 for five- and seven-point scales, and *AD* should not exceed 2.0 for 11-point scales. Finally, it should be noted that Brown and Hauenstein (2005) proposed rules of thumb for $a_{wg}$. Specifically, 0–0.59 was considered unacceptable, 0.60–0.69 was weak, and 0.70–0.79 was moderate, and above 0.80 was strong agreement.

## Statistical Standards

Identifying standards for IRA using statistical significance testing involves conducting Monte Carlo simulations or random group resampling. For Monte Carlo simulations, the input is the correlation matrix of scale items, the null distribution, and the significance level (see Cohen et al., 2001, 2009; Burke and Dunlap, 2002; Dunlap et al., 2003). Tabled significance values were provided by several researchers (e.g., Dunlap et al., 2003; Cohen et al., 2009). The program *R* contains commands for running Monte Carlo simulations involving $r_{wg}$ and *AD* (see Bliese, 2009). The objective is to create a sampling distribution for the IRA statistic with an expected mean and standard deviation, which can be used to generate confidence intervals and significance tests. Random group resampling involves constructing a sampling distribution by repeatedly sampling and forming random groups from observations in the observed data set, and comparing the significance of the mean difference in within-group variances of the observed distribution and the randomly generated distribution using a *Z*-test (Bliese et al., 2000; Bliese and Halverson, 2002; Ludtke and Robitzsch, 2009), for which commands are available in *R*. Thus, significance testing of the $S_{wg}$ is possible through the random-group resampling approach. Similar logic could be applied to test the significance of $r^*_{wg}$, $a_{wg}$, and $CV_{wg}$, although existing scripts for running these tests may be more difficult to find.

Statistical significance testing of IRA statistics has its advantages. Cutoff criteria are relatively objective, thereby potentially reducing misuse by relying on inappropriate or arbitrary rules of thumb (see below). But, statistical significance does not appear to have been widely implemented. One reason might be because of the novelty of the methods for doing so, and the need to understand and implement commands in *R*, for example. Another reason might be because statistical agreement might be difficult to reach in many commonly-encountered practical situations. Specifically, many applications will involve three to five raters, yet $r_{wg(j)}$ needs to be in the range of at least 0.75 and AD would have to fall below 0.40 (Burke and Dunlap, 2002; Cohen et al., 2009) in order to reject the null hypothesis of no agreement. Indeed, Cohen et al. reported that groups with low sample sizes rarely reached levels of statistical significance that would allow the hypothesis of no statistical agreement to be rejected. If statistical significance testing is treated as a hurdle against which agreement must be passed in order for further consideration of the implicated variables, there

is potential to interfere with advancement of research involving low (but typical) sample sizes. This may not always be the most desirable application of IRA, and, not surprisingly, practical standards have tended to be most common.

## Current Best Practice in Judging Agreement Levels

It is important to acknowledge the two divergent purposes of practical and statistical approaches to judging agreement. Practical cutoffs provide decision rules about whether or not agreement seems to have exceeded a minimum threshold in order to justify a decision. Examples of such decisions include aggregation of lower-level data to higher-level units, retention of critical incidents in job analyses, and for assessing whether frame-of-reference training has successfully "calibrated" raters. The use of practical cutoff criteria in these decisions implies that a certain level of agreement is needed in order to make some practical decision in light of the agreement qualities of the data (Burke and Dunlap, 2002).

Statistical standards are not focused on the absolute level of agreement so much as they are concerned with drawing inferences about a population given a sample. Statistical agreement tests the likelihood that the observed agreement in the sample is greater than what would be expected by chance at a certain probability value (e.g., $p < 0.05$). It involves making inferences about whether the sample was most likely drawn from a population with chance levels of agreement vs. systematic agreement. For example, a set of judges could be asked to rate the job relevance of a personality variable in personality oriented job analysis (Goffin et al., 2011). If agreement is not significant for a particular variable, it would suggest that there is no systematic agreement in the population of judges (Cohen et al., 2009). Notice that this differs from practical significance, which would posit a cutoff, above which agreement levels would be considered adequate for supporting the use of the mean rating as an assessment of the job relevance of the trait (e.g., O'Neill et al., 2011).

Statistical agreement raises issues of power and sample size. Specifically, in small samples statistical agreement will be more difficult to reach than in large samples. Accordingly, outcomes of whether agreement is strong or not may depend on whether one focuses on statistical or practical decision standards, and in large samples, statistical agreement alone should not sufficiently justify aggregation (Cohen et al., 2009). The key point, however, is that statistical significance testing is for determining whether the agreement level for a particular set of judges exceeds chance levels. Practical agreement is about absolute levels of agreement in a sample, which could be seen as strong even for non-significant agreement when sample sizes are low.

In light of the above discussion, it is clear that more research is needed in order to identify defensible and practical approaches for judging IRA levels. Best practice recommendations for the interim would involve reporting several IRA statistics, ideally from different families, in order to provide a balanced perspective on IRA. Practical significance levels could be advanced *a priori* using suggestions described above (e.g., Burke and Dunlap, 2002; Brown and Hauenstein, 2005; LeBreton and Senter, 2008) in order to identify cutoffs for making decisions. Statistical significance would be employed only when inferences about the population are important and when a power analysis suggests sufficient power to detect agreement, although practical standards should also be considered especially when power is very high. Thus, a researcher or practitioner might place little emphasis on statistical significance when he or she is not concerned about generalizing to the population, and when there are very few judges the researcher might be advised to consider a less stringent significance level (e.g., $\alpha = 0.10$). Importantly, when evaluating agreement in a set of judges, the focus is typically not on whether the sample was drawn from a population with chance or systematic agreement, but whether there is a certain practically meaningful level of agreement. Thus, in many cases practical significance might be most critical.

Interpretations of practical agreement should probably not be threshold-based, all-or-none decision rules applied to a single statistic [e.g., satisfactory vs. unsatisfactory $r_{wg(j)}$]. This is how statistics can be misused to support a decision (see Biemann et al., 2012). Rather, reporting the values from several IRA statistics along with proposed practical standards of agreement reviewed here will provide some evidence of the quality of the ratings, which can be considered in the context of other important indices that also reflect data quality (e.g., reliability, validity). An overall judgment can then be advanced and the reader (including the reviewer) will also have the necessary information upon which to form his or her own judgment. This procedure fits well within the spirit of the unitary perspective on validity (Messick, 1991; Guion, 1998), which suggests that validity involves an expert judgment on the basis of all the available reliability and validity evidence regarding a construct. It would seem that IRA levels should be considered in the development of this judgment, but it may not be productive to always require an arbitrary level of agreement to support or disconfirm the validity of a measure in a single study. In any case, consequential validity (Messick, 1998, 2000) should be kept in mind, and more research examining the consequences, implications, and meaning of various standards for IRA is needed.

# CONCLUSION

IRA statistics are critical to justification of aggregation in multilevel research, but they are also frequently applied in job analysis, performance appraisal, assessment centers, employment interviews, and so forth. Importantly, IRA offers a unique perspective from reliability because reliability deals with consistency of ratings and agreement deals with the similarity of absolute levels of ratings. IRA has the added advantage of providing one estimate per set of raters—not one estimate for the sample as is the case with reliability. This feature of IRA can be helpful for diagnostic purposes, such as identifying particular groups with high or low IRA.

Despite the prevalence of IRA, there is the problem that articles considering IRA statistics tend to be heavy on the technicals (e.g., Lindell and Brandt, 1999; Cohen et al., 2001), and this might be a reason why $r_{wg}$, with its widely known limitations (e.g., see Brown and Hauenstein, 2005), appears to persevere as the leading statistical choice for IRA. Indeed, a recent review suggested that a lack of a sound understanding

of IRA statistics may have led to some misuses (see Biemann et al., 2012). Thus, despite the many alternatives offered (e.g., $r^*_{wg}$, $AD$, $a_{wg}$, $CV_{wg}$), they may not receive full consideration because accessible, tractable, and non-technical resources describing each within a framework that allows for simple contrasting is not available. LeBreton and Senter (2008) provided solid coverage, but it was mainly with respect to multilevel aggregation issues and not directly applicable to other purposes (e.g., job analysis).

The current article aims to fill a gap in earlier research by offering an introductory source, intended to be useful for scholars with a wide range of backgrounds, in order to facilitate application and interpretation of IRA statistics. Through a comparative analysis regarding eight IRA statistics, it appears that these statistics are not interchangeable and that they are differentially affected by various contextual details (e.g., number of Likert response options, number of judges, number of scale items). The goal of the article is to facilitate critical and appropriate applications of IRA in the future, offer a foundation for tackling the more technical sources currently available, and make suggestions regarding best practices in light of the insights gleaned through the review. It is proposed that researchers interpret IRA levels with respect to the situation and best-practice recommendations for practical and statistical standards in the literature, as reviewed here. Because of the unique limitations of each statistic, it is probably safe to conclude that more than one statistic should always be reported. In submissions where this has been ignored, reviewers should request the author to report additional agreement statistics, ideally from other IRA families. Consistent with the unitary perspective on validity, it is suggested that judgments regarding the adequacy of the ratings rely on evidence of IRA in conjunction with additional statistics that shed light on the quality of the data (e.g., reliability coefficients, criterion validity coefficients). Regarding agreement standards, it would seem advisable to evaluate a given IRA statistic using appropriate *a priori* practical cutoffs and statistical criteria, depending on the purpose of assessing agreement levels. What we need to avoid is misuses of agreement statistics and adoption of inappropriate or misleading decision rules. This critical review aims to provide tools to help researchers and practitioners avoid these problems.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

## ACKNOWLEDGMENTS

This research was supported by an operating grant provided to TO by the Social Sciences and Humanities Research Council of Canada.

## REFERENCES

Allen, N. J., Stanley, D. J., Williams, H. M., and Ross, S. J. (2007). Assessing the impact of nonresponse on work group diversity effects. *Organ. Res. Methods* 10, 262–286. doi: 10.1177/1094428106/294731

Allison, P. D. (1978). Measures of inequality. *Am. Sociol. Rev.* 43, 865–880. doi: 10.2307/2094626

Bedeian, A. G., and Mossholder, K. W. (2000). On the use of the coefficient of variation as a measure of diversity. *Organ. Res. Methods* 3, 285–297. doi: 10.1177/109442810033005

Bernardin, H. J., and Walter, C. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *J. Appl. Psychol.* 62, 64–69. doi: 10.1037/0021-9010.62.1.64

Biemann, T., Cole, M., and Voelpel, S. (2012). Within-group agreement: on the use (and misuse) of rWG and rWG(J) in leadership research and some best practice guidelines. *Leadersh. Q.* 23, 66–80. doi: 10.1016/j.leaqua.2011.11.006

Bliese, P. D. (2009). *Multilevel Modeling in R (2.3): A Brief Introduction to R, the Multilevel Package and Nlme Package*.

Bliese, P. D., and Halverson, R. R. (2002). Using random group resampling in multilevel research. *Leadersh. Q.* 13, 53–68. doi: 10.1016/S1048-9843(01)00104-7

Bliese, P. D., Halverson, R. R., and Rothberg, J. (2000). *Using Random Group Resampling (RGR) to Estimate Within-Group Agreement with Examples Using the Statistical Language R*. Walter Reed Army Institute of Research.

Brown, R. D., and Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: an alternative to the rwg indices. *Organ. Res. Methods* 8, 165–184. doi: 10.1177/1094428105275376

Brutus, S., Fleenor, J. W., and London, M. (1998). Does 360-degree feedback work win different industries? A between-industry comparison of the reliability and validity of multi-source performance ratings. *J. Manage. Dev.* 17, 177–190. doi: 10.1108/EUM0000000004487

Burke, M. J., and Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: a user's guide. *Organ. Res. Methods* 5, 159–172. doi: 10.1177/1094428102005002002

Burke, M. J., Finkelstein, L. M., and Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organ. Res. Methods* 2, 49–68. doi: 10.1177/109442819921004

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions: a comparison of intraclass correlation coefficients, rwg(j), hierarchical linear modeling, within- and between-analysis, and random group resampling. *Leadersh. Q.* 13, 69–93. doi: 10.1016/S1048-9843(01)00105-9

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: a typology of composition models. *J. Appl. Psychol.* 82, 234–246. doi: 10.1037/0021-9010.83.2.234

Cohen, A., Doveh, E., and Eick, U. (2001). Statistical properties of the rwg(j) index of agreement. *Psychol. Methods* 6, 297–310. doi: 10.1037/1082-989X.6.3.297

Cohen, A., Doveh, E., and Nahum-Shani, I. (2009). Testing agreement for multi-item scales with the indices $r_{wg(j)}$ and $AD_{m(j)}$. *Organ. Res. Methods* 12, 148–164. doi: 10.1177/1094428107300365

Crawley, M. J. (2007). *The R Book*. Chichester: W*iley*.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York, NY: Wiley.

Dunlap, W. P., Burke, M. J., and Smith-Crowe, K. (2003). Accurate tests of statistical significance for rwg and average deviation interrater agreement indices. *J. Appl. Psychol.* 88, 356–362. doi: 10.1037/0021-9010.88.2.356

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educ. Psychol. Meas.* 30, 71–76. doi: 10.1177/001316447003000106

Flanagan, J. C. (1954). The critical incident technique. *Psychol. Bull.* 327–358. doi: 10.1037/h0061470

George, J. M. (1990). Personality, affect, and behavior in groups. *J. Appl. Psychol.* 86, 1075–1082. doi: 10.1037/0021-9010.75.2.107

George, J. M., and James, L. R. (1993). Personality, affect, and behavior in groups revisited: comment on aggregation, levels of analysis, and a recent application of within and between analysis. *J. Appl. Psychol.* 78, 798–804. doi: 10.1037/0021-9010.78.5.798

Goffin, R. D., Rothstein, M. G., Reider, M. J., Poole, A., Krajewski, H. T., Powell, D. M., et al. (2011). Choosing job-related personality traits: developing

valid personality-oriented job analysis. *Pers. Individ. Dif.* 51, 646–651. doi: 10.1016/j.paid.2011.06.001

Guion, R. M. (1998). *Assessment, Measurement, and Prediction for Personnel Decisions*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Harvey, R. J. (1991). "Job analysis," in *Handbook of Industrial and Organizational Psychology*, Vol. 2, eds M. D. Dunnette and L. M. Hough (Palo Alto, CA: Consulting Psychologists Press), 71–163.

Harvey, R. J., and Hollander, E. (2004, April). "Benchmarking rWG interrater agreement indices: let's drop the.70 rule-of-thumb," in *Paper Presented at the Meeting of the Society for Industrial and Organizational Psychology* (Chicago, IL).

James, L. R., Demaree, R. G., and Wolf, G. (1984). Estimating within group interrater reliability with and without response bias. *J. Appl. Psychol.* 69, 85–98. doi: 10.1037/0021-9010.69.1.85

James, L. R., Demaree, R. G., and Wolf, G. (1993). rwg: an assessment of within group interrater agreement. *J. Appl. Psychol.* 78, 306–309. doi: 10.1037/0021-9010.78.2.306

Klein, K. J., Conn, A. B., Smith, D. B., and Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *J. Appl. Psychol.* 86, 3–16. doi: 10.1037/0021-9010.86.1.3

Kozlowski, S. W. J., and Hattrup, K. (1992). A disagreement about within-group agreement: disentangling issues of consistency versus consensus. *J. Appl. Psychol.* 77, 161–167. doi: 10.1037/0021-9010.77.2.161

Kozlowski, S. W. J., and Hults, B. M. (1987). An exploration of climates for technical updating and performance. *Pers. Psychol.* 40, 539–563. doi: 10.1111/j.1744-6570.1987.tb00614.x

Kozlowski, S. W. J., and Klein, K. J. (2000). "A multilevel approach to theory and research in organizations: contextual, temporal, and emergent processes," in *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions,* eds K. J. Klein and S. W. J. Kozlowski (San Francisco, CA: Jossey Bass), 3–90.

Lance, C. E., Butts, M. M., and Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: what did they really say? *Organ. Res. Methods* 9, 202–220. doi: 10.1177/1094428105284919

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., and James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: are ratings from multiple sources really dissimilar? *Organ. Res. Methods* 6, 78–126. doi: 10.1177/1094428102239427

LeBreton, J. M., James, L. R., and Lindell, M. K. (2005). Recent issues regarding rWG, $r^*_{WG}$, $r_{WG(J)}$, and $r^*_{WG(J)}$. *Organ. Res. Methods* 8, 128–138. doi: 10.1177/1094428104272181

LeBreton, J. M., and Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organ. Res. Methods* 11, 815–852. doi: 10.1177/1094428106296642

Lindell, M. K. (2001). Assessing and testing interrater agreement on a single target using multi-item rating scales. *Appl. Psychol. Meas.* 25, 89–99. doi: 10.1177/01466216010251007

Lindell, M. K., and Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Appl. Psychol. Meas.* 21, 271–278. doi: 10.1177/01466216970213006

Lindell, M. K., and Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: a comparison of the CVI, T, $r_{wg(j)}$, and $r^*_{wg(j)}$ Indexes. *J. Appl. Psychol.* 84, 640–647. doi: 10.1037/0021-9010.84.4.640

Lindell, M. K., and Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *J. Appl. Psychol.* 85, 331–348. doi: 10.1037/0021-9010.85.3.331

Lindell, M. K., Brandt, C. J., and Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Appl. Psychol. Meas.* 23, 127–135. doi: 10.1177/01466219922031257

Ludtke, O., and Robitzsch, A. (2009). Assessing within-group agreement: a critical examination of a random-group resampling approach. *Organ. Res. Methods* 461–487. doi: 10.1177/1094428108317406

Meade, A. W., and Eby, L. T. (2007). Using indices of group agreement in multilevel construct validation. *Organ. Res. Methods* 10, 75–96. doi: 10.1177/1094428106289390

Messick, S. (1991). *Validity of Test Interpretation and Use*. Research Report for the Educational Testing Service, Princeton, NJ.

Messick, S. (1998). Test validity: a matter of consequences. *Soc. Indic. Res.* 45, 35–44. doi: 10.1023/A:1006964925094

Messick, S. (2000). "Consequences of test interpretation and test use: the fusion of validity and values in psychological assessment," in *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy*, eds R. D. Goffin and E. Helmes (Norwell, MA: Klewer Academic Publishers), 3–20.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. New York, NY: McGraw-Hill.

Morgeson, F. P., and Campion, M. A. (2000). Accuracy in job analysis: toward an inference-based model. *J. Organ. Behav.* 21, 819–827. doi: 10.1002/1099-1379(200011)21:7<819::AID-JOB29>3.0.CO;2-I

Newman, D. A., and Sin, H.-P. (2009). How do missing data bias estimates of within-group agreement? Sensitivity of $SD_{wg}$, $CV_{wg}$, $r_{wg(j)}$, $r_{wg(j)^*}$, and ICC to systematic nonresponse. *Organ. Res. Methods* 12, 113–147. doi: 10.1177/1094428106298969

Nunnally, J. C. (1978). *Psychometric Theory, 2nd Edn.* New York, NY: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory 3rd Edn.* New York, NY: McGraw-Hill.

O'Neill, T. A., and Allen, N. J. (2014). Team task conflict resolution: an examination of its linkages to team personality composition and team effectiveness outcomes. *Group Dyn.* 18, 159–173. doi: 10.1037/gdn0000004

O'Neill, T. A., Lewis, R. J., and Carswell, J. J. (2011). Employee personality, justice perceptions, and the prediction of workplace deviance. *Pers. Individ. Dif.* 51, 595–600. doi: 10.1016/j.paid.2011.05.025

O'Neill, T. A., McLarnon, M. J. W., and Carswell, J. J. (2015). Variance components of job performance ratings. *Hum. Perform.* 28, 66–91. doi: 10.1080/08959285.2014.974756

Pasisz, D. J., and Hurtz, G. M. (2009). Testing for between-group differences in within-group interrater agreement. *Organ. Res. Methods* 12, 590–613. doi: 10.1177/1094428108319128

Roberson, Q. M., Sturman, M. C., and Simons, T. L. (2007). Does the measure of dispersion matter in multilevel research? A comparison of the relative performance of dispersion indexes. *Organ. Res. Methods* 10, 564–588. doi: 10.1177/1094428106294746

Rousseau, D. M. (1985). "Issues of level in organizational research: multilevel and cross level perspectives," in *Research in Organizational Behavior,* Vol. 7, eds L. Cummings and B. Saw (Greenwich, CT: JAI Press), 1–37.

Schmidt, A. M., and DeShon, R. P. (2003, April). "Problems in the use of rwg for assessing interrater agreement," in *Paper Presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology* (Orlando, FL).

Schmidt, F. L., and Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *J. Appl. Psychol.* 74, 368–370. doi: 10.1037/0021-9010.74.2.368

Schneider, B., Salvaggio, A. N., and Subirats, M. (2002). Climate strength: a new direction for climate research. *J. Appl. Psychol.* 87, 220–229. doi: 10.1037/0021-9010.87.2.220

van Mierlo, H., Vermunt, J. K., and Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organ. Res. Methods* 12, 368–392. doi: 10.1177/1094428107309322

Walker, A. G., and Smither, J. W. (1999). A five year study of upward feedback: what managers do with their results matters. *Pers. Psychol.* 52, 393–423. doi: 10.1111/j.1744-6570.1999.tb00166.x

Weingart, L. R., Olekalns, M., and Smith, P. L. (2004). Quantitative coding of negotiation behavior. *Int. Negot.* 9, 441–456. doi: 10.1163/1571806053498805

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.