Statistical Significance Criteria for the $r_{WG}$ and Average Deviation Interrater Agreement Indices

Kristin Smith-Crowe  
University of Utah

Michael J. Burke  
Tulane University

Ayala Cohen  
Technion – Israel Institute of Technology

Etti Doveh  
Technion – Israel Institute of Technology


**Acknowledgments:** This research was supported by funding from the David Eccles School of Business at the University of Utah.

*This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.*

http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=2013-35326-001
Abstract

Despite the widespread use of interrater agreement statistics for multilevel modeling and other types of research, the existing guidelines for inferring the statistical significance of interrater agreement are quite limited. They are largely relevant only under conditions that numerous researchers have argued rarely exist. Here we address this problem by generating guidelines for inferring statistical significance under a number of different conditions via a computer simulation. As a set, these guidelines cover many of the conditions researchers commonly face. We discuss how researchers can use the guidelines presented to more reasonably infer the statistical significance of interrater agreement relative to using the limited guidelines available in the extant literature.

Keywords: interrater agreement, multilevel research, $r_{WG}$, average deviation, AD, aggregation, levels of analysis, variance
Statistical Significance Criteria for the $r_{WG}$ and Average Deviation Interrater Agreement Indices

Numerous problems in applied psychology and management call for assessing interrater agreement, such as employees’ ratings of the work climate within an organization (e.g., Chuang & Liao, 2010), the classification of studies into a meta-analytic category (e.g., Burke, Chan-Serafin, Salvador, Smith, & Sarpy, 2008), subject matter experts’ perspectives on the use of a selection procedure (e.g., Murphy, Cronin, & Tam, 2003), raters’ assessments of items for measurement development purposes (e.g., Bledow & Frese, 2009), customers’ ratings of satisfaction or loyalty with respect to a store (e.g., Salanova, Agut, & Peiro, 2005), and coders’ classification of qualitative data into a category (e.g., Smith-Crowe, Burke, & Landis, 2003) to name just a few. In particular, assessments of interrater agreement have become increasingly important in multilevel modeling studies over the last decade, where researchers make determinations of whether data measured at the individual level of analysis can be aggregated for use at a higher (e.g., workgroup or business unit) level of analysis. In fact, within the applied psychology literature alone there has been a strong and steady increase over the last decade in the use of interrater agreement indices to justify data aggregation in multilevel modeling research on organizational climate (e.g., see Chen, Lam, & Zhong, 2007; Colquitt, Noe, & Jackson, 2002; Dawson, Gonzalez-Roma, Davis, & West, 2008; Ehrhart, 2004; Gonzalez-Roma, Fortes-Ferreria, & Peiro, 2009; Hoffman & Mark, 2006; Klein, Conn, Smith, & Sorra, 2001; Neal & Griffin, 2006; Simons & Roberson, 2005; Takeuchi, Chen, & Lepak, 2009; Wallace, Popp, & Mondore, 2006).

Notably, $r_{WG}$ (James, Demaree, & Wolf, 1984), which represents the observed variance in ratings compared to the variance of a theoretical distribution representing no agreement (i.e., the null distribution), and the Average Deviation (AD; Burke, Finkelstein, & Dusig, 1999), which
represents the average deviations in ratings from the mean or median rating, are commonly employed by researchers to address interrater agreement problems including whether or not data can be aggregated from lower to higher levels of analysis in multilevel modeling studies. Yet, despite the widespread use of interrater agreement indices, such as $r_{WG}$ and AD, controversy has arisen over the appropriate criteria to be used for assessing agreement (Biemann, Cole, & Voelpel, 2012; Brown & Hauenstein, 2005; LeBreton & Senter, 2008; Meyer, Mumford, & Campion, 2010). In large part, this controversy reflects the fact that criteria for the statistical significance of interrater agreement estimates have only been developed for a very limited set of circumstances. That is, as discussed in more detail below, statistical significance criteria exist only for single items with respect to only one null distribution, the uniform distribution (Burke & Dunlap, 2002; Smith-Crowe & Burke, 2003; Dunlap, Burke, & Smith-Crowe, 2003), or for scales with respect to only two null distributions, the uniform and slightly skewed distributions (Cohen, Doveh, & Nahum-Shani, 2009).

The current paper addresses this gap in the literature by developing an extensive set of criteria for assessing the statistical significance of two commonly employed interrater agreement indices: $r_{WG}$ and AD. The criteria we present can be used across a broad array of research problems in applied psychology and management where different forms of responding or response bias operate; we provide examples of when these different criteria would be relevant. Importantly, these criteria will allow researchers to make more informed inferences with regard to data aggregation in multilevel modeling studies. Below, we elaborate on the interrater agreement indices, $r_{WG}$ and AD, including how they have been used and interpreted in the extant literature. Next, we describe the limitations of the current criteria for assessing the adequacy of interrater agreement, arguing for the need for additional criteria relative to additional null
response distributions and discussing when they would be relevant. Finally, we describe our objective of generating critical values for assessing the statistical significance of interrater agreement estimates relative to a number of possible null distributions.

**r\textsubscript{WG} and Average Deviation (AD) Indices**

Although a number of procedures and indices for assessing interrater agreement have been proposed (see LeBreton & Senter, 2008 for a review), we focus on two commonly used indices of interrater agreement: r\textsubscript{WG} and the Average Deviation (AD). r\textsubscript{WG} and AD focus on the degree to which raters’ assessments of a single target are interchangeable. For the purpose of data aggregation, this point is nontrivial as within-group agreement to justify data aggregation is not conditional on between group differences (George & James, 1993). Furthermore, while r\textsubscript{WG} is the more frequently employed index, we also include AD due to its increased usage in the applied psychology and management literatures, as well as its being singled out for performing well relative to other interrater agreement indices in simulation research (Roberson, Sturman, & Simons, 2007).

As proposed by James et al. (1984), the r\textsubscript{WG} index is a comparison between the observed variance in ratings and the variance of a null distribution (i.e., a theoretically specified distribution representing no agreement). Demonstrating agreement, then, is a matter of showing that the variance of the observed ratings is sufficiently less than the variance that would be expected under conditions of no agreement. Specifically, r\textsubscript{WG} for a single item (r\textsubscript{WG(j)}) is calculated as follows:

\[
r\textsubscript{WG(j)} = 1 - \left( \frac{s^2}{\sigma^2} \right),
\]

where \(s^2\) is the observed variance in ratings of item \(j\) and \(s^2\) is the variance of the null distribution. Because the ratio of observed to theoretically specified variance is subtracted from
1, higher \( r_{WG} \) values indicate greater agreement. \( r_{WG} \) for a scale, or multiple, parallel items is calculated as

\[
r_{WG(J)} = \frac{J[1-(s^2/\sigma^2)]}{J(1-(s^2/\sigma^2)) + s^2/\sigma^2},
\]

where \( J \) is the number of items, \( s^2 \) is the mean observed variance in ratings of \( J \) items, and other notations are the same as for Equation 1.

Importantly, for the calculation of \( r_{WG} \), it is necessary to specify a null distribution representing no agreement so that the observed variance can be compared to the null variance. In other words, researchers must make a decision about what value to plug in for \( \sigma^2 \). Researchers almost exclusively define the null distribution as the uniform, or rectangular distribution, where each response category is equally likely. The variance for the uniform distribution is calculated as

\[
\sigma_{\bar{y}}^2 = (A^2 - 1)/12,
\]

where \( A \) is the number of response categories. Use of the uniform distribution as the null distribution assumes that no response bias is present and that each response option on a Likert-type scale is equally likely (James et al., 1984; LeBreton & Senter, 2008).

Use of the uniform distribution as the null distribution, or the denominator in the calculation of \( r_{WG} \), has been routinely criticized as being often inappropriate (Brown & Hauenstein, 2005; Meyer et al., 2010) because the assumption that no response bias is present is likely untenable in many situations. Inappropriate usage of the uniform distribution is problematic because the choice of the null distribution is key to assessing and interpreting agreement (Biemann et al., 2012; Brown & Hauenstein, 2005; Cohen et al., 2009; James et al., 1984; LeBreton & Senter, 2008; Lüdtke & Robitzsch, 2009; Meyer et al., 2010). Further, as
LeBreton and Senter explained, using the uniform distribution as the null distribution is likely to inflate $r_{WG}$ values because alternative null distributions (e.g., skewed, triangular, and normal distributions) have lower variances, and when lower values are plugged into the denominator of the $r_{WG}$ formula, resulting $r_{WG}$ values are lower, indicating less agreement. In other words, by using the uniform distribution as the null distribution when other distributions would be theoretically more appropriate, researchers are drawing inferences and making decisions to aggregate data to higher levels of analysis based on inflated $r_{WG}$ values. As we will discuss in the next section, few guidelines exist for interpreting $r_{WG}$ values in terms of statistical significance when null distributions other than the uniform distribution are used. Thus, while researchers have been admonished to more thoughtfully choose a null distribution, realistically they only have limited opportunities to do so without an expansion of the criteria for determining statistical significance.

Burke and colleagues (1999) proposed the Average Deviation (AD) index as a more straightforward method for quantifying agreement, one that does not require the specification of a null distribution in order to assess within-group agreement. AD is calculated as the average deviation of ratings for an item from the mean ($AD_M$) or median ($AD_{Md}$) rating for an item:

\[
AD_M(j) = \frac{\sum_{k=1}^{N} |x_{jk} - \bar{x}_j|}{N},
\]

\[
AD_{Md}(j) = \frac{\sum_{k=1}^{N} |x_{jk} - Md_j|}{N},
\]

where $N$ is the number of raters, $x_{jk}$ is the $k^{th}$ judge’s score on item $j$, $\bar{x}_j$ is the mean rating for item $j$, and $Md_j$ is the median rating for item $j$. Because AD is a measure of dispersion, lower numbers indicate higher agreement. AD can also be calculated for a scale, or multiple, parallel
items. $AD_M(j)$ and $AD_{Md}(j)$ are calculated as the average deviations for J items from the mean ($AD_M(j)$) or median ($AD_{Md}(j)$), respectively.

Though AD does not require the specification of a null distribution in order to compute agreement, in order to interpret AD values, one must compare them to some notion of no agreement (i.e., a null distribution). As such, like $r_{WG}$, AD requires researchers to theoretically define no agreement. Further, once researchers do choose how they will define no agreement (i.e., which null distribution they will use), they need guidelines for drawing inferences about agreement based on their comparison of observed agreement to the null distribution, or no agreement. As is the case with $r_{WG}$, few guidelines exist for interpreting AD in terms of statistical significance when null distributions other than the uniform distribution are used.

**Limitations of the Existing Criteria for Assessing Interrater Agreement**

Often researchers wish to make claims about the statistical significance of interrater agreement, or whether the observed agreement is due to chance. In this section we discuss limitations in existing guidelines available to researchers for interpreting AD and $r_{WG}$ values in terms of statistical significance. Further, we identify a number of null distributions for which there are no statistical significance guidelines, and we discuss when these distributions would be relevant.

**Null distributions for which criteria exist.** Burke and Dunlap (2002) first published statistical significance guidelines for $AD_M(j)$ via an approximate randomization test. If observed values are less than or equal to the tabled critical values, then researchers can state that the observed values are statistically significant ($p<.05$). Using the same methodology, Dunlap et al. (2003) published similar guidelines for $r_{WG}(j)$. Researchers with observed values equal to or
greater than the tabled values can state that their observed values are statistically significant (p<.05).

Though useful, both sets of critical values for statistical significance (those of Burke and Dunlap, 2002, and those of Dunlap et al., 2003) are limited in two important ways. First, they apply to agreement for individual items only, not scales. Second, they are based exclusively on the uniform distribution as the null distribution. It is worth noting that the uniform distribution is sometimes the right distribution. When no response bias is expected, for instance, the uniform distribution would be the appropriate distribution (e.g., James et al., 1984). The uniform distribution may also be appropriate when raters face conceptual ambiguity such as rating more ambiguous white-collar work, compared to more straightforward blue-collar work (Smith-Crowe, Burke, Kouchaki, & Signal, 2013). Yet, the uniform distribution is often not appropriate as the null distribution (e.g., Lebreton & Senter, 2008).

Using a simulation methodology similar to that of Burke and Dunlap (2002) and Dunlap et al. (2003), Cohen et al. (2009; see also Cohen, Doveh, & Eick, 2001) addressed both of these limitations by generating critical values for \( r_{WG(J)} \) and \( AD_{M(J)} \) (i.e., agreement on 3 and 5-item scales) relative to a slightly skewed null distribution, as well as the uniform null distribution. Observed \( r_{WG(J)} \) values equal to or greater than the tabled values and observed \( AD_{M(J)} \) values equal to or less than the tabled values are statistically significant (p<.05).

While Cohen et al.’s (2009) paper expanded upon the previously available criteria for inferring the statistical significance of observed interrater agreement values, the available guidelines remain limited. For single items, the existing statistical significance criteria are only relevant when the uniform distribution is the appropriate null distribution. For scales, the existing statistical significance criteria are only relevant when the uniform or slightly skewed
distributions are the appropriate null distributions. Yet, as we discuss next, numerous other distributions are relevant to a great deal of applied psychology and management research.

**Null distributions for which criteria are needed.** In order to address the problem of limited guidelines for assessing the statistical significance of interrater agreement, we must identify null distributions likely to be broadly relevant to researchers, but for which there are currently no guidelines for inferring statistical significance. To do so, we drew on previous discussions of the relevance of different null distributions, particularly those discussions presented by James et al. (1984), LeBreton and Senter (2008), and Smith-Crowe et al. (2013). We identify skewed distributions, triangular and normal distributions, and bimodal and subgroup distributions as those for which criteria are needed, and we articulate the circumstances under which they would be relevant. We summarize our discussion of the relevance of different null distributions in Table 1 so as to provide a convenient guide for researchers who are considering which null distributions might be appropriate to their research questions.

First, criteria for assessing statistical significance relative to null distributions representing greater than slight skew (i.e., moderately skewed and heavily skewed distributions) are needed. Skewed distributions are appropriate null distributions when factors such as social desirability, leniency, or severity biases are expected (James et al., 1984; LeBreton & Senter, 2008; Smith-Crowe et al., 2013). For instance, supervisors may report subordinates’ performance positively out of a desire to be supportive or to look good themselves (James et al.); subordinates may show leniency bias when rating their supervisors (Ng, Koh, Ang, Kennedy, & Chan, 2011; Schriesheim, 1981) or coworkers (Wang, Wong, & Kwong, 2010); and the social
desirability of items can restrict the range of responses (Klein et al., 2001). Skew may also be a product of the verbal anchors used for Likert-type scales. French-Lazovik and Gibson (1984) found that unusually positive anchors resulted in a shift toward the negative end of the scale, and unusually negative anchors resulted in a shift toward the more positive end of the scale.

Arguably when multiple factors known to lead to skewed response distributions are potentially operating, then researchers would benefit from an assessment of interrater agreement relative to a null distribution with moderate to high skew. For instance, Ng et al. (2011) found that subordinates’ performance ratings of superiors were more lenient than peer or supervisor ratings of the same focal person, and that the degree of leniency was even greater for subordinates who more strongly valued power distance and collectivism. These results suggest that to the extent performance ratings are given by subordinates who value power distance and collectivism, null distributions representing greater skew should be employed.

Second, criteria for assessing statistical significance relative to triangular and normal null distributions are needed. James et al. (1984) argued that the appropriate null distribution would be a triangular or normal distribution when central tendency bias is expected (see also LeBreton & Senter, 2008; Smith-Crowe et al., 2013), such as when inexperienced or untrained raters are asked to respond to complex or vague items, or when raters feel compelled to be cautious and respond to items using the middle of the scale (e.g., for political reasons). For instance, Bernardin and Buckley (1981) describe how police sergeants’ ratings of subordinates were highly leptokurtic after an authoritative endorsement of bell-shaped rating distributions by a high ranking police official.

Third, criteria for assessing statistical significance relative to bimodal and subgroup distributions are needed. Schwarz (1999) discussed how features of a self-report instrument,
such as the nature of the response scale (e.g., a low-frequency scale ranging from “never” to “once a month” versus a high-frequency scale ranging from “twice a month or less” to “several times a day”), coupled with the degree to which a behavior is represented in memory can force respondents to rely on estimation strategies (e.g., using the scale itself as a frame of reference) and lead to systematic bias in judgments. Notably, Schwarz’s work would suggest the consideration of bimodal or subgroup null response distributions when raters are asked to provide retrospective behavioral reports on frequency scales. Here, a subgroup response distribution refers to the situation where responses tend to cluster around two response options and in unequal proportions. A number of other authors provide rationales for why one might expect subgroups in regard to responding to questions about polarized issues and beliefs (e.g., see Nicklin & Roch, 2009; Harrison & Klein, 2007). Lindell and Brandt (1997) suggested, for instance, that when raters of research proposals come from different academic disciplines, a split along disciplinary lines might be expected.

It is important to note here that multiple null distributions may be relevant, and, thus, multiple null distributions should be used in these cases. For instance, Rego, Vitória, Magalhães, Ribeiro, and Cunha (2013) present this discussion of their rationale to use both uniform and slightly skewed distributions in assessing interrater agreement in their study:

For computing the expected variances that allow calculating $r_{WG(j)}$ values, several authors (e.g., Biemann et al., 2012; LeBreton & Senter, 2008) recommend using several defensible null distributions. Thus, expected variances of team virtuousness, AL [authentic leadership], and team affective commitment were estimated assuming both a uniform (rectangular) null distribution (“the most natural candidate to represent nonagreement”; Cohen, Doveh, & Nahum-Shani, 2009, p. 149) and a slightly skewed
distribution. We considered slightly skewed distribution based on earlier studies where measures of perceptions of organizational virtuousness (Rego et al., 2010, 2011), authentic leadership (Rego et al., 2012a), and affective commitment (Rego et al., 2011) were included. For team potency, we also considered it reasonable to expect a slightly skewed distribution, because of a possible leniency bias on the part of the followers when describing their teams. (p. 72)

Rego et al.’s (2013) thoughtfulness regarding what null distributions they should use is in line with the advice given by James et al. (1984) in their seminal paper on interrater agreement:

“Ask the question: If there is no true variance in the judgments and the true [interrater agreement] is zero, then what form of distribution would be expected to result from response bias, and, of course, some random measurement error? This distribution reflects one’s hypothesis about response bias and is referred to as the ‘null distribution.’ (If no systematic bias is expected, then the null distribution is uniform.)” (p. 90). After almost 25 years of James et al.’s advice being largely disregarded or overlooked by researchers, LeBreton and Senter (2008) restated this advice as a strongly worded injunction: “Given the ubiquity of response biases in organizational research, we call for a moratorium on the unconditional (i.e., unjustified) use of any null distribution, especially the uniform distribution. Instead, we challenge each researcher to justify the use of a particular null distribution, uniform or otherwise, in his or her research study” (p. 830). We agree fully with James et al. and LeBreton and Senter, yet we also sympathize with the researcher, who like Rego et al. would like to be thoughtful in the selection of relevant null distributions, but who has few options of null distributions for which there are relevant guidelines for assessing interrater agreement.

**Objectives of the Current Study**
To address the deficiencies in regard to assessments of interrater agreement, a primary objective of this investigation is to generate critical values assessing the statistical significance of $r_{WG}$ and AD values relative to skewed, triangular, normal, bimodal, subgroup, and uniform null response distributions. In doing so, we present critical values both for single items and scales, and with respect to different numbers of raters, response categories, items per scale, and correlations among items within a scale. Though the vast majority of the critical values presented herein have not been published before, we do present values for the uniform null distribution (for items and scales) and a slightly skewed null distribution (for scales), which have been presented in previous work (Burke & Dunlap, 2002; Cohen et al., 2009; Dunlap et al., 2003). We regenerated critical values for these null distributions so as to provide a more complete set of critical values in the current paper. In what follows, we describe our simulation methodology and the results of the simulations.

**Method**

A computer simulation program was written in R programming language (2011), using Orddata library for generating correlated artificial ordinal data (Leisch, Weingessel, Kaiser, & Hornik, 2011). The software (in R) and details on the simulations are available online at http://ie.technion.ac.il/Labs/Stat/. It is consistent with previous programs written to generate statistical significance cut-off values for interrater agreement statistics (Cohen et al., 2001; Cohen et al., 2009; Dunlap et al., 2003).

We should note, however, that our methodology is different from the random group resampling (RGR) methodology which also exploits the availability of computer power to

---

1 The software will enable obtaining critical values for parameters not included in the current paper. More specifically, researchers can input different null distributions (which implicitly specifies the number of response scale options), group sizes, either numbers of items and degrees of inter-item correlation (or a user specified inter-item correlation matrix), and numbers of iterations in the simulation.
The RGR method compares within group variances of actual groups to within group variances of pseudo groups. Significant agreement is inferred from the variance of actual groups being less than that of pseudo groups. In contrast, our methodology is based on the original notion of assessing interrater agreement via $r_{WG}$ and $AD$: observed data are compared to an a priori, theoretically relevant null distribution representing no agreement. To understand the advantage of our methodology, consider an extreme example where there is agreement among all persons surveyed regardless of group membership (e.g., everyone responds with a 4 or a 5 on a 5-point Likert scale). In this case the within group variances of the actual groups will be roughly equal to that of the pseudo groups, indicating a lack of agreement according to the RGR method, when there is in fact a high level of agreement. Because our methodology entails assessing agreement based on a null distribution, it is not vulnerable to a failure to detect significant agreement in a case like this one.

For the purposes of the current paper, the program input included the sample size ($N$, or the number of raters), the number of items per scale, the number of response options, the correlation between items of a scale ($\rho$), and the theoretical null response distribution. Alpha was set to .05. Critical values were computed for agreement on items ($AD_{M(j)}$, $AD_{Md(j)}$, $r_{WG(j)}$) and scales ($AD_{M(J)}$, $AD_{Md(J)}$, $r_{WG(J)}$). Rationales for the specific values of these inputs are discussed below.

The null response distributions included in the simulations are shown in Table 2 in terms of the proportion endorsing each response category. These distributions were based on those used by other researchers (Burke et al., 1999; James et al., 1984; LeBreton & Senter, 2008; Messick, 1982; Smith-Crowe et al., 2013). Note that for the bimodal and subgroup distributions,
“moderate” and “extreme” refer to how far apart on the scale raters are. For the subgroup distributions, “A” and “B” refer to how large the minority is. Importantly, the null response distributions in Table 2 arguably apply to the vast majority of interrater agreement problems in applied psychology and management insofar as responding along 4-point, 5-point, and 7-point Likert-type response scales are concerned.

A representative set of sample sizes (N=5, 10, 30, 100) was chosen to reflect small, moderate, and relatively large groups of respondents. These values represent variations in sample sizes for workgroups, teams, and survey respondents found in studies concerning interrater agreement problems. For instance, authors such as Chuang and Liao (2010) and Towler, Lezotte, and Burke (2011) report average sample sizes per store for a variety of retail contexts of approximately 10 individuals per store. In measurement development studies and qualitative analyses/studies concerned with interrater agreement, authors, such as Bledow and Frese (2009) among many others, report sample sizes in the range of 25 to 30. Finally, in sample survey studies where authors are examining issues of consensus vs. controversy among groups of subject matter experts, larger samples (typically of 100 or greater) are reported (e.g., see Murphy et al., 2003; Nicklin & Roch, 2009).

The numbers of items per scale (3, 5, 10), as well as the number of response options (4-point, 5-point, 7-point response scales), were also chosen to reflect typical ranges for these measurement characteristics. That is, there are numerous examples in the literature, particularly that relevant to data aggregation, of 3, 5, and 10 item scales (e.g., Borucki & Burke, 1999; Hoffman, Gorman, Blair, Meriac, Overstreet, & Atchley, 2012; Katz-Navon, Naveh, & Stern,
 Likewise, there are numerous examples of 4-point, 5-point, and 7-point response scales (Borucki & Burke, 1999; Dawson et al., 2008; Katz-Navon, Naveh, & Stern, 2009; Scandura & Graen, 1984; Tordero, González-Romá, & Peiró, 2008; Towler et al., 2011).

Finally, in choosing the values for inter-item correlations (rhos) we examined both theoretical and empirical literatures related to theoretical limits, and we subsequently chose the following values: .2, .3, .4, .5, .6. In determining the theoretical range for rho, we relied in part on a fundamental and lesser known equation in the development of the theory of measurement error that specifies the relationship between the average correlation among items, test length, and reliability (i.e., Equation 6-18 of Nunnally, 1967). This equation can be written as follows:

$$r_{kk} = \frac{k \times \text{Avg } r_{ii}}{1 + (k-1) \text{Avg } r_{ii}}$$

(6)

where $r_{kk}$ is the reliability for a $k$ item measure, and Avg $r_{ii}$ is the average correlation among the $k$ items.

Using this equation, reasonable or likely theoretical limits for the average correlation among items within a measure can be established for 3, 5, and 10 item measures. Average inter-item correlations that range between .2 and .6 would produce test reliabilities that respectively range between .38 and .82 for 3-item measures, .56 and .88 for 5-item measures, and .71 and .94 for 10-item measures. To verify this choice of values for rho in our simulations, we checked the vast literature on interrater agreement and worked with the above formula and the reported internal consistency reliabilities to solve for the average inter-item correlation. For instance, Sowinski, Fortmann, and Lezotte (2008) aggregated data for two 5-item work climate scales with reported reliabilities of .77 and .80. The above formula indicates that the average inter-item correlations for these scales would be approximately .40 to .44, values which are well within the
theoretical limits we have chosen for rho. The same point would hold for almost all other studies that we examined.

**Results and Discussion**

**Critical Values for Single Item Measures**

The critical values for $AD_{M(j)}$, $AD_{Md(j)}$, and $r_{WG(j)}$ (i.e., agreement relative to single items) are shown in Table 3.\(^2\) The values are listed for combinations of number of raters ($N=5, 10, 30, 100$), response scale categories ($4, 5, 7$), and null response distribution. Observed values of $AD_{M(j)}$ and $AD_{Md(j)}$ equal to or less than the corresponding tabled values are statistically significant ($p \leq .05$). Observed values of $r_{WG(j)}$ equal to or greater than the corresponding tabled values are statistically significant ($p \leq .05$). Below, we discuss how to use the tabled values with respect to examples drawn from the literature.

\[\text{Insert Table 3 here.}\]

To give an example of how interrater agreement estimates and the critical values in Table 3 for statistical significance can be applied, we consider data reported by Nicklin and Roch (2009) concerning experts’ perspectives on letters of recommendation (LOR). For one item that Nicklin and Roch considered representative of “consensus” among experts (“I believe letter inflation is a problem with LORs that will never be entirely alleviated”), 90% of the experts rated the item as agree or strongly agree on a 5-point response scale and 10% of the ratings were in the “neutral” or “disagree” response options. The affective (attitudinal) nature of this item coupled with the nature of the responding might suggest that the null response distribution reflect

\(^2\) Here we should note that in some cases of small sample size, the critical values listed in Table 3 represent perfect agreement ($r_{WG}=1, AD=0$); these values also occur in Tables 4-6, which are discussed in the next section. In these cases, there is not sufficient evidence to be able to reject the null hypothesis. Yet, it does not mean that we can conclude there is no agreement.
moderate to heavy skew. The observed $AD_{M(j)}$ value was .57. If a moderately skewed distribution were used as the null distribution, the observed agreement would be interpreted as statistically significant because the observed value for AD of .57 is lower than the critical value for the moderately skewed distribution ($N=100$) of .62. However, using a heavily skewed distribution as the theoretical null response distribution, the observed agreement would be interpreted as statistically nonsignificant because the observed value for AD of .57 is greater than the critical value for a heavily skewed distribution ($N=100$) of .55.

As a second example with summary data from Murphy et al. (2003), we consider an item and topic, the relative merits of cognitive ability testing, that is known to reflect subgroup beliefs (see Rynes, Colbert, & Brown, 2002): “General cognitive ability enhances performance in all domains of work.” For this item and with respect to responses on a 5-point response scale with anchors ranging from strongly disagree to strongly agree, 30% responded in the “disagree” range and 62% responded in the “agree” range. If the null response distribution were a moderate subgroup (B), then the observed $AD_{M(j)}$ of .85 would be statistically nonsignificant as it is above the critical value of .28 for an N of 100. However, the observed AD value is significant relative to a uniform null response distribution (where the critical value is 1.08 for sample size of 100). This example illustrates the potential value of examining interrater agreement relative to theoretical null response distributions other than exclusively relying on a uniform null response distribution.

Returning to the summary data from Nicklin and Roch (2009), let’s consider an item that they classified as “polarized” (i.e., “I have written a LOR for an individual I did not know very well”). For this item, 49% of the respondents agreed and 45% disagreed, with approximately 6% using the “neutral” response option. The nature of this item in terms of representing different
beliefs about helping behavior might suggest the consideration of a bimodal or subgroup null response distribution. Nicklin and Roch reported an observed AD value 1.13 for this item on a 5-point response scale. If the null response distribution were a moderate bimodal distribution, then there is no statistically significant agreement at the .05 level with respect to $AD_{M(j)}$ and a sample size of 100 as the observed AD is greater than the critical value of .96, but there is statistically significant agreement if the null response distribution is the uniform distribution. These results also point to the value of examining interrater agreement in a potentially more theoretically appropriate manner with respect to null response distributions other than the uniform null distribution.

This point holds for usage of $r_{WG(j)}$ as well. For example, in their network study on voice, Venkataramani and Tangirala (2010) assessed coworkers’ agreement on a single item measuring specific employees’ personal influence: “How influential do you think this person is in your branch?” One hundred ninety coworkers in 42 groups responded to this item using a 5-point scale (1=not at all influential, 5=extremely influential). The authors calculated $r_{WG(j)}$ using the uniform distribution as the null distribution, resulting in a value of .82; the authors concluded that there was sufficient agreement to warrant including this measure in their analyses. Indeed, compared against the uniform distribution as the null distribution and assuming a group size of 5, this value is statistically significant (see Table 3). Rather than using the uniform distribution as the null distribution or the only null distribution, however, a skewed distribution is likely an appropriate null distribution. As we noted previously, leniency (Ng et al., 2011; Wang et al., 2010) and the social desirability of items (Klein et al., 2001) would be expected to skew ratings. In this case, coworkers rated each other regarding a socially desirable characteristic (i.e.,
influence). Given that multiple factors promoting skew are in place, using a moderately or heavily skewed distribution as the null distribution would be reasonable.

Importantly, then, how would the use of these alternative distributions potentially affect the inferences one can draw about the degree of agreement in the Venkataramani and Tangirala (2010) study? In order to address this question, we recalculated the $r_{WG(j)}$ for this study. We used Equation 8 in the Appendix, as well as the information provided by the authors for $r_{WG(j)}$ and the variance of the null distribution (.82 and 2, respectively) to solve for the observed variance ($s^2$). The resulting value for $s^2$ is .36. With this value for the observed variance, we calculated $r_{WG(j)}$ using moderately and heavily skewed distributions as the null distributions. Based on the proportions listed in Table 1, we calculated the variances for moderately skewed and heavily skewed distributions: $\sigma = .91$ and .44, respectively. Using Equation 1 to recalculate $r_{WG(j)}$ based on these alternative null distributions, we arrived at these values for $r_{WG(j)}$: .60 (moderately skewed distribution) and .18 (heavily skewed distribution). Referring to Table 3, the critical values for $r_{WG(j)}$ for the moderately and heavily skewed null distributions ($N=5$) are .78 and .55, respectively. Thus, relative to a moderately or heavily skewed null distribution, one cannot infer that agreement is statistically significant. This example is another example suggesting the importance of the distribution chosen to represent no agreement.

**Critical Values for Multi-Item Measures**

The critical values for $AD_{M(J)}$, $AD_{Md(J)}$, and $r_{WG(J)}$ for 3-item scales are shown in Table 4, those for 5-item scales are shown in Table 5, and those for 10-item scales are shown in Table 6. The critical values in Tables 4-6 are listed for combinations of number of raters ($N=5, 10, 30, 100$), correlation among items ($r=.20, .30, .40, .50, .60$), response scale categories (4, 5, 7), and null distribution. Observed values of $AD_{M(J)}$ and $AD_{Md(J)}$ equal to or less than the corresponding
tabled values are statistically significant \((p \leq 0.05)\). Observed values of \(r_{WG(J)}\) equal to or greater than the corresponding tabled values are statistically significant \((p \leq 0.05)\).

_____________________

Insert Tables 4-6 here.

_____________________

To give an example of how interrater agreement estimates for measurement scales and the critical values in Tables 4-6 for statistical significance can be applied, we consider summary statistics reported in Towler et al. (2011) with respect to employees’ work climate assessments. Towler et al. used a 5-item measure of work climate where items were rated on a 5-point scale from strongly disagree to strongly agree. For an average sample size per store of 10 employees, these authors reported an average \(AD_{M(J)}\) value of 0.55. Assuming an average correlation among items of 0.3 and using a slightly skewed null response distribution, each of their observed \(AD_{M(J)}\) values would have had to be less than 0.71 to be statistically significant (see Table 5). A cautionary note, however, is that the tabled critical values are not intended to apply to average interrater agreement statistics, but are proposed for use with respect to each target in a study (e.g., each store in the Towler et al. study).

As another example, consider Salanova, Agut, and Peiró’s (2005) study where they reported interrater agreement findings for work engagement scales. Among Salanova et al.’s scales they measured dedication on a 5-item scale with items scored on a 7-point frequency rating scale. For each work unit (restaurant or hotel) they collected data from 3 to 10 employees and reported an average \(AD\) value of 0.22. Assuming an average correlation of 0.4 among items on the dedication scale, 10 employees per unit, and a moderately skewed null response distribution, an observed \(AD\) below the critical value of 0.89 (Table 5), would be considered significant.
Given the range of AD values reported in Salanova et al., from .15 to .31, each of the observed AD values for the 114 units in their study was below this critical value.

As a final example, we consider Farh, Lee, and Farh’s (2010) study of task conflict and team creativity. To measure task conflict, they used a 4-item measure. Team members from 71 project teams responded to the items using a 5-point scale. The authors reported a median $r_{WG(J)}$ value of .91; they used the uniform distribution as the null distribution. As with the Towler et al. (2011) study, it is important to note here that the critical values presented in the tables are meant to be used for individual agreement statistics (in this case, the $r_{WG(J)}$ calculated for each team) rather for means or medians of agreement statistics across all groups sampled. For illustration purposes, however, we discuss here the median $r_{WG(J)}$ value reported by Fahr et al. Comparing their reported value to Table 5, we see that it is statistically significant ($\alpha=.05; N=5; \rho=.30; \text{null distribution}=\text{uniform}$), as the critical value of .75 is less than .91. Similarly, the corresponding critical value reported in Table 4 for a 3-item scale of .72 is less than .91.

It would be reasonable, however, to use a skewed null distribution in this case. Farh et al. (2010) collected their data at a Chinese multi-national company. Based on Ng et al.’s (2011) work we might expect the leniency bias to be a factor in these data, as those who are higher in collectivism show greater leniency in peer ratings than those lower in collectivism; in this case, participants higher in collectivism may be reluctant to negatively rate the team on task conflict. Further, the social desirability of items can skew ratings (Klein et al., 2001); here, it may be socially desirable to deflate ratings of team task conflict. Using Equation 15 reported in the Appendix, we can solve for the observed variance ($s^2$) in team members’ ratings of task conflict using the information reported by Farh et al. to fill in the other values in the equation ($J=4$, $r_{WG(J)}=.91$, $\sigma^2=2$). The resulting value of $s^2$ was .57. Based on the proportions listed in Table 2,
we calculated the variance for a moderately skewed distribution: \( \sigma = .91 \). Using Equation 2, we find that using a moderately skewed distribution as the null distribution, the resulting value of \( r_{WG(J)} \) is .71. This value is not statistically significant because it is lower than the critical value of .82 reported in Table 5. It is also lower than the corresponding critical value reported in Table 4: .79. Here again we see the importance of the choice of null distribution in inferring the statistical significance of interrater agreement.

**General Discussion**

The critical values presented in this article offer alternatives to the heavy reliance on the uniform null response distribution when studying interrater agreement in applied psychology and management. Arguably, the use of null response distributions that take into account theoretical and methodological bases for response distributions other than the uniform distribution will serve to advance theory testing in regard to interrater agreement problems. Although hypothesis testing relative to the uniform null response distribution can be useful for some interrater agreement problems, most interrater agreement problems call for the consideration of alternative null response distributions that reflect different forms of response bias under the null. We advocate for the use and reporting of multiple interrater agreement indices and null response distributions in assessments of interrater agreement. Over time, we would expect refinements in theory development, as well as empirical findings, to provide more explicit guides for researchers’ beliefs about the appropriateness of particular null response distributions in an area of inquiry. In the long run, triangulating on the appropriate null response distributions within particular areas of inquiry and giving consideration to multiple null response distributions within individual studies will lead to not only more informed judgments about interrater agreement, but also to greater precision in theory.
Importantly, though, there are a number of questions surrounding the statistical significance of interrater agreement that remain to be addressed. For instance, how can researchers efficiently use agreement indices to make inferences concerning the extent of agreement across multiple groups? Often, such agreement is demonstrated by quoting the median or the mean of the collection of indices corresponding to the upper-level units (e.g., the median or the mean of the observed agreement index across organizations, or units within organizations). Though the median or the mean of $r_{WG}$, $AD_M$, or $AD_{Md}$ obtained across upper-level units may indicate high agreement, there often is variability among the upper-level units in their level of agreement that is not captured by simply quoting the median or the mean. Cohen et al. (2009) briefly discussed methods for drawing inferences concerning whether the level of agreement is different from “no-agreement” on the basis of the ensemble of upper-level units. Future research could extend our work and that of Cohen et al. (2009) by developing inference methods for assessing the extent of agreement, while accounting for variability in agreement across upper-level units, as well as for size differences across upper-level units.

The above comments should not be construed as a call for a moratorium or movement away from the use of practical assessments and standards for interrater agreement. Practical assessments are informative with respect to assessing the level of agreement in a set of ratings, making determinations about the quality of aggregated data as indicators of constructs at higher levels of analysis, and with respect to assessing within-group agreement for intact groups or where data are census in nature. However, practical significance critical values (see Burke & Dunlap, 2002; Smith-Crowe et al., 2013) can be used in combination with statistical significance critical values for making assessments of interrater agreement when sampling has occurred. Practical assessments of interrater agreement offer an interpretive angle concerning the
magnitude of interrater agreement, or the meaningfulness of the agreement, which along with
tests of statistical significance in regard to the use of non-uniform null response distributions
may serve to better inform inferences about interrater agreement.

In conclusion, we presented critical values for significance tests that can be applied to a
wide variety of interrater agreement problems and theoretical null response distributions in
applied psychology and management. We also provided an extensive discussion, including
numerous examples of when these different null distributions would be relevant. In doing so, we
have extended the use of interrater agreement assessments at the item level as well as the scale
level. These critical values are intended to assist researchers in making more informed
judgments about interrater agreement with respect to the aggregation of individual-level data to
higher levels of analysis and in regard to other types of interrater agreement problems.
References


Smith-Crowe, K., & Burke, M. J. (2003). Interpreting the statistical significance of observed AD interrater agreement values: Correction to Burke and Dunlap (2002). *Organizational Research Methods, 6*, 127-129.


Table 1

Relevance of Null Response Distributions

<table>
<thead>
<tr>
<th>Null Distribution</th>
<th>Conditions for Relevance</th>
<th>Literature-Based Examples of When Distributions May be Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewed</td>
<td>• Social desirability</td>
<td>• Subordinate ratings of supervisor performance, especially if subordinates value collectivism and power distance (Ng et al., 2011)</td>
</tr>
<tr>
<td></td>
<td>• Leniency bias</td>
<td>• Measurement of socially desirable constructs (Klein et al., 2001)</td>
</tr>
<tr>
<td></td>
<td>• Severity bias</td>
<td>• Use of unusually positive or negative verbal anchors for Likert-type scales (French-Lazovik &amp; Gibson, 1984)</td>
</tr>
<tr>
<td></td>
<td>• Overly positive or negative anchor valence</td>
<td></td>
</tr>
<tr>
<td>Triangular and</td>
<td>• Central tendency bias</td>
<td>• Dictated adherence to a bell-shaped distribution (Bernardin &amp; Buckley, 1981)</td>
</tr>
<tr>
<td>Normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bimodal and</td>
<td>• Retrospective reports of behavior</td>
<td>• Use of retrospective reports of behavioral frequency (Schwarz, 1999)</td>
</tr>
<tr>
<td>Subgroup</td>
<td>• Factions</td>
<td>• Measurement of polarized beliefs or regarding polarized issues (Nicklin &amp; Roch, 2009; Harrison &amp; Klein, 2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Surveys of subgroups (Lindell &amp; Brandt, 1997)</td>
</tr>
<tr>
<td>Uniform</td>
<td>• Absence of response bias</td>
<td>• Measurement that entails conceptual ambiguity (Smith-Crowe et al., 2013)</td>
</tr>
<tr>
<td></td>
<td>• Conceptual ambiguity</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* The conditions for relevance listed here are not exhaustive, but are instead meant to be representative and to include a number of typical situations researchers face. The literature-based examples also are not exhaustive. Furthermore, researchers must carefully consider their own unique circumstances before attempting to generalize from these examples.