Assessing Interrater Agreement via the Average Deviation Index

Given a Variety of Theoretical and Methodological Problems

Kristin Smith-Crowe
University of Utah
David Eccles School of Business
1655 East Campus Center Drive
Salt Lake City, Utah 84112
kristin.smith-crowe@business.utah.edu

Michael J. Burke
Tulane University
Freeman School of Business
New Orleans, Louisiana 70118
mburke1@tulane.edu

Maryam Kouchaki
University of Utah
David Eccles School of Business
1655 East Campus Center Drive
Salt Lake City, Utah 84112
maryam.kouchaki@business.utah.edu

Sloane Signal
College of Education and Human Development
Jackson State University
1400 J.R. Lynch Street, JSU Box 17209
Jackson, Mississippi 39217-0209
Sloane.m.signal@students.jsums.edu

Smith-Crowe, K., Burke, M. J., Kouchaki, M., & Signal, S. (2013). Assessing interrater agreement via the average deviation index given a variety of theoretical and methodological problems. *Organizational Research Methods 16,* 127-151.

Author Note

Kristin Smith-Crowe, David Eccles School of Business, University of Utah; Michael J. Burke, Freeman School of Business, Tulane University; Maryam Kouchaki, David Eccles School of Business, University of Utah; Sloane Signal, Jackson State University.

We would like to thank Greg Oldham and Isaac Smith for helpful comments on previous drafts of the paper, and Julie Seidel and Teng Zhang for research assistance.

Correspondence concerning this article should be addressed to Kristin Smith-Crowe, David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, 801-587-3720 (phone). E-mail: kristin.smith-crowe@business.utah.edu.

Abstract

Currently, guidelines do not exist for applying interrater agreement indices to the vast majority of methodological and theoretical problems that organizational and applied psychology researchers encounter. For a variety of methodological problems, we present critical values for interpreting the practical significance of observed average deviation (AD) values relative to either single items or scales. For a variety of theoretical problems, we present null ranges for AD values, relative to either single items or scales, to be used for determining whether an observed distribution of responses within a group is consistent with a theoretically specified distribution of responses. Our discussion focuses on important ways to extend the usage of interrater agreement indices beyond problems relating to the aggregation of individual level data.

Assessing Interrater Agreement via the Average Deviation Index Given a Variety of Theoretical and Methodological Problems

Assessments of interrater agreement, or the degree to which raters are interchangeable (Kozlowski & Hattrup, 1992)[1], are integral to many types of organizational and applied psychology research. For instance, interrater agreement assessments have recently been central with respect to addressing substantive questions within domains such as organizational climate and leadership (e.g., Dawson, Gonzalez-Roma, Davis, & West, 2008; Walumbwa & Schaubroeck, 2009), conducting quantitative and qualitative research, as well as laboratory and field studies (e.g., Katz-Navon, Naveh, & Stern, 2009; Kreiner, Hollensbe, & Sheep, 2009; Van Kleef, Homan, Beersma, Knippenberg, Knippenberg, & Damen, 2009), developing measures (e.g., Bledow & Frese, 2009; Lawrence, Lenk, & Quinn, 2009), dealing with various types of data analysis problems (e.g., Grant & Mayer, 2009; Nicklin & Roch, 2009; Trougakos, Beal, Green, & Weiss, 2008), and deciding whether or not to aggregate data (e.g., Borucki & Burke, 1999; Takeuchi, Chen, & Lepak, 2009). Further, usage of interrater agreement statistics is on the rise. In the *Journal of Applied Psychology* and *Personnel Psychology* alone, there has been a largely linear increase in the use of these statistics over the past decade (see Figure 1). Notably, in 2010 almost half of the articles published in these journals used interrater agreement statistics.

Despite the relevance of interrater agreement assessments for dealing with a broad array of theoretical and methodological issues and their widespread usage, systematically derived guidelines for applying interrater agreement indices to the vast majority of problems that researchers and practitioners encounter do not exist. The primary objective of this paper is to

[1] Interrater agreement is distinct from interrater reliability (e.g., Wagner, Rau, & Lindermann, 2010). While the former is concerned with the extent to which ratings are the same across raters, the latter is concerned with consistency in the rank order of ratings across raters. Kozlowski and Hattrup (1992) helpfully distinguished "consensus" (agreement) from "consistency" (reliability).

derive practical guidelines to assist researchers using the average deviation (AD) index in making more informed decisions about interrater agreement problems. We focus on the AD index, the average deviation from the mean or median of ratings, for two primary reasons. First, AD is straightforward. It measures agreement, while intraclass correlations (ICC) measure both agreement and reliability simultaneously (LeBreton & Senter, 2008), potentially complicating inferences. Further, for both ICC and r_{WG} researchers must choose from among numerous variations to employ the statistic (see LeBreton & Senter, 2008, for a review). Second, AD performs well. In a simulation study Roberson, Sturman, and Simons (2007) found that the AD index performs as well as similar other statistics. Kline and Hambley (2007) reported similar findings.

Importantly, we are concerned with practical significance, or "…whether an index indicates that interrater agreement is sufficiently strong or disagreement is sufficiently weak so that one can trust that the average opinion of a group is interpretable or representative…"[2] (Dunlap, Burke, & Smith-Crowe, 2003, p. 356), as practical significance is the basis on which agreement is typically evaluated. We present critical values for addressing the frequently asked methodological question concerning practical significance, "How much agreement/dispersion is there?" These critical values can be used to assess agreement on a single item or a scale. This question concerns the *level* of agreement in a set of ratings. An answer to this question often informs decisions about the quality of a measure of central tendency, such as a group's mean, as an indicator of the group's standing on a phenomenon or construct of interest. While previous work has also addressed this question, as we will discuss below, the guidelines provided are of very limited use.

[2] While practical significance concerns whether agreement is meaningful, statistical significance concerns whether it is due to chance (Dunlap et al., 2003; Smith-Crowe & Burke, 2003).

In particular, we go beyond the work of Burke and Dunlap (2002), who previously provided a decision rule for interpreting the practical significance of observed AD values, to provide decision rules that cover many more circumstances. As we detail below, though the calculation of AD does not require the specification of a null distribution representing no agreement, the interpretation of AD does. In other words, while one can calculate AD in the absence of a specified null distribution, one cannot draw conclusions regarding observed AD values without comparing them to some notion of "no agreement." Burke and Dunlap's guideline is based exclusively on the uniform distribution as the null distribution; there are no guidelines for interpreting the practical significance of AD relative to any other null distributions. Below we discuss the criticisms of researchers' overreliance on the uniform distribution despite other distributions often being more appropriate. Herein, we provide guidelines for interpreting AD in terms of the level of agreement relative to numerous other distributions. Our guidelines will allow researchers to interpret interrater agreement relative to null distributions more appropriate to their research than the uniform distribution.

Furthermore, we present guidelines for addressing the less commonly posed, yet theoretically important question of "How well does the pattern of observed agreement/dispersion match the theoretically specified pattern of agreement/dispersion?" These guidelines can be used in relation to either agreement on a single item or a scale. An answer to this question informs decisions regarding the scoring of the group as consistent or not with the theoretically specified distribution and, thus, the use of such scores in subsequent analyses at the group level of analysis. Addressing questions related to the *pattern* of dispersion will be of increasing importance as researchers attempt to test new theories concerning group and other higher level phenomena which specify patterns of dispersion as variables (e.g., see DeRue, Hollenbeck, Ilgen,

& Feltz, 2010; Harrison & Klein, 2007). By focusing on the pattern in addition to level of agreement/dispersion, our work promotes conceptual advances in research and goes beyond previous work on interrater agreement (e.g., Burke & Dunlap, 2002).

For the purpose of demonstrating how our guidelines would be used to address problems relating to the pattern of dispersion, we will focus on notions of diversity and team efficacy dispersion, as theories relating to these phenomena have recently been presented. For the purpose of demonstrating how our guidelines would be applied to the assessment of the level of agreement, we focus our discussion on the common use of interrater agreement indices for data aggregation decisions. The guidelines we present, however, would apply to the study of a broad array of interrater agreement problems.

To unfold our discussion, we begin with a brief summary of research on multilevel modeling and data aggregation to set the stage for a discussion related to assessments of the level of agreement. This discussion also includes an overview of the relevance of interrater agreement assessments for determining whether or not the observed pattern of dispersion matches a theoretically specified pattern of dispersion. Then, we present interpretive standards for assessments of interrater agreement for both the level of agreement and pattern of dispersion, with detailed discussions of how the derived guidelines can be applied to a variety of research problems.

Issues Related to the Level of Agreement and Pattern of Dispersion

In this section we discuss the use of interrater agreement in multilevel research to justify the aggregation of lower level data to higher levels of analysis based on the level of observed agreement. Then, we discuss a second possible usage of interrater agreement statistics, which is to assess the goodness of fit between an observed pattern of dispersion with a theoretically

specified pattern of dispersion. We give examples of recent multilevel theories that predict outcomes based on patterns of dispersion. Related to both level of agreement and pattern of dispersion, we discuss the limited availability of guidelines available to researchers for interpreting agreement.

Level of Agreement

Multilevel research commonly entails researchers aggregating data so as to create measures or indicators of higher level constructs. The appropriateness of representing higher level constructs by aggregating individual level data is established by a composition model, which represents theory on how multilevel constructs are related at each level of analysis (Chan, 1998; Kozlowski & Klein, 2000; see also Klein, Dansereau, & Hall, 1994; Rousseau, 1985). For instance, Chan's (1998, p. 236) direct consensus model is the idea that the "meaning of [the] higher level construct is in the consensus among lower levels"; the referent-shift consensus model is the idea that the "lower level units being composed by consensus are conceptually distinct though derived from the original individual-level units"; and the dispersion model is the idea that the "meaning of [the] higher level construct is in the dispersion or variance among lower level units." Importantly, composition arguments indicate the type of evidence needed to justify the aggregation of individual level data, with several models, including the direct consensus and referent-shift models (Chan, 1998), specifying interrater agreement, or the interchangeability of raters, as the appropriate type of evidence. Interrater agreement is also important for dispersion models (Chan, 1998); in this case, the degree of agreement itself represents the higher level construct.

Essentially, interrater agreement via the average deviation (AD) index is established by demonstrating that observed agreement is sufficiently greater than no agreement. Thus, though it

is not necessary to the calculation of AD, in order to assess, or interpret, observed AD values, researchers must identify an appropriate random response distribution, or null distribution, to which observed variability in responses can be compared. A number of scholars have cited the choice of a null distribution as key to interpreting indices of interrater agreement, and thus drawing appropriate inferences from data (e.g., Brown & Hauenstein, 2005; Cohen, Doveh, & Nahum-Shani, 2009; James, Demaree, & Wolf, 1984; LeBreton & Senter, 2008; Lindell & Brandt, 1997; Lüdtke & Robitzsch, 2009; Meyer, Mumford, & Campion, 2010). In practice, however, researchers routinely rely on the uniform distribution as the null distribution, though doing so is likely often inappropriate (e.g., Brown & Hauenstein, 2005; Meyer et al., 2010). In fact, LeBreton and Senter (2008) recently called for a moratorium on the unconditional reliance on the uniform distribution.

The consequences of inappropriately comparing observed data to the uniform null distribution can be (a) that researchers mistakenly do not read interrater agreement as being sufficient for aggregation to higher levels of analysis, (b) that researchers mistakenly read interrater agreement as being sufficient for aggregation to higher levels of analysis (e.g., see Meyer et al., 2010), or (c) that researchers fail to appropriately interpret a group's standing on a variable of interest. Thus, comparing observed data to an inappropriate null distribution can lead to erroneous inferences that have important implications for researchers. Nonetheless, the only decision rule for interpreting the practical significance of observed AD values is Burke and Dunlap's (2002) decision rule, which is based on the uniform distribution as the null distribution. Currently, there are no guidelines for interpreting practical significance relative to any other distributions.

While assessments of within group agreement for methodological purposes, such as data aggregation as discussed above, address the question, "How much agreement/dispersion is there?", another question researchers can answer using interrater agreement indices is "How well does the pattern of observed agreement/dispersion match the theoretically specified pattern of agreement/dispersion?" Below we discuss the issue of the pattern of dispersion and the theoretical distributions to which observed patterns can be compared.

Pattern of Dispersion

Harrison and Klein (2007) recently argued for the theoretical import of considering the pattern of dispersion. They distinguished among separation diversity (e.g., differences in opinions, beliefs, or attitudes), variety diversity (e.g., differences in knowledge or experience), and disparity diversity (e.g., differences in proportionate ownership or control over socially valued assets). They argued that depending on the type of diversity, minimum, moderate, and maximum diversity would be associated with differently shaped distributions; that is, both the type and degree of diversity determine the shapes of distributions. For instance, maximum separation diversity is characterized by a bimodal distribution, maximum variety diversity is characterized by a uniform distribution, and maximum disparity diversity is characterized by a skewed distribution. For separation diversity, minimum, moderate, and maximum degrees of diversity are characterized as unimodal, uniform, and bimodal, respectively. Considering both type of diversity and pattern of dispersion, they argued that maximum separation diversity (bimodal distribution) and maximum disparity diversity (skewed distribution) will have negative outcomes, such as reduced cohesion and group member input, respectively, while maximum variety diversity (uniform distribution) will have positive outcomes, such as increased creativity.

Importantly, according to their theory, both the type of diversity and the pattern of dispersion must be known in order to effectively predict outcomes. For example, separation diversity could be measured with regard to team members' opinions about what their teams' goals are (Harrison & Klein, 2007). For each team, the pattern of the distribution of these opinions would be compared to unimodal, uniform, and bimodal distributions as these are the distributions theoretically specified by Harrison and Klein as representing minimum, moderate, and maximum separation diversity. The degree of separation diversity, then, would be indicated by the theoretical distribution that is most similar to the observed distribution. With this measure of degree of separation diversity for each team, in addition to measures of cohesion, conflict, trust, and performance, researchers could test Harrison and Klein's (2007) hypothesis that as the degree of separation diversity increases, team outcomes will be more negative: less cohesion and trust, more conflict, and lower performance.

DeRue et al.'s (2010) work on team efficacy provides another example of the potential theoretical importance of the *pattern* of dispersion above and beyond the *level* of dispersion. They argued that teams could have the same level of dispersion in their team efficacy ratings, but have different theoretically meaningful patterns of dispersion. These different patterns of dispersion, they argued, would predict different outcomes. Thus, according to DeRue et al.'s (2010) theory of team efficacy dispersion, assessing the pattern of dispersion in team efficacy ratings is essential for making predictions about team effectiveness. For instance, they argued that while a bimodal distribution of team efficacy ratings would lead to both positive and negative outcomes, a uniform distribution would lead to positive outcomes. Regarding the effects of a uniform distribution, their argument was that their disagreement will lead team members' to share their differing views, thus enhancing team structuring, planning, and learning,

while simultaneously allowing the team to avoid problems of extreme magnitudes of efficacy, which can lead either to over-confidence or helplessness, and social factions, which create dysfunctional conflict. In contrast, they argue that a bimodal distribution will similarly lead to team members' sharing their differing views and thus enhancing team processes, but due to the existence of social factions, will also lead to dysfunctional conflict.

While the question of the *level* of dispersion has been important for various reasons, especially justifying aggregating individual level data to form higher level variables, it is likely that the question of the *pattern* of dispersion will become increasingly important as more researchers consider the theoretical import of response distributions in and of themselves. This forecast is consistent with a recent call from Edwards and Berry (2010) to increase the theoretical precision in management research by developing hypotheses that specify effects in terms of magnitude, form (linear, nonlinear, etc.), and conditions (i.e., moderators). In reviewing 25 years (1985-2009) of articles published in the *Academy of Management Review*, Edwards and Berry (2010) found that 10.4% of the propositions stated only that a relationship would exist, and 89.6% only indicated the direction of the relationship. The theories presented by DeRue et al. (2010) and Harrison and Klein (2007) are important steps toward more precise management theories because they consider the shapes of distributions rather than simply measures of central tendency.

In cases for which the pattern of dispersion is of interest, it will be necessary to specify a "null response range," analogous to a null range with regard to a formal test of the null hypothesis (see Greenwald, 1975), to determine whether the observed pattern of responses, or the relative percentages of individuals within the respective categories, is consistent with the theoretical distribution. To date, though researchers have suggested that observed patterns of

dispersion can be quantitatively assessed (DeRue et al., 2010; Harrison & Klein, 2007), no one has developed practical guidelines for drawing inferences about the goodness-of-fit between an observed distribution and a theoretically specified distribution. As such, practical guidelines are needed for addressing both the methodological question of the level of agreement/dispersion and the theoretical question of the pattern of responses.

Summary

In order to address this dearth of guidelines, we specify a variety of response distributions that researchers could use to address a number of theoretical and methodological issues, and we derive decision rules for the AD index relevant to each of these distributions to aid researchers in making inferences about interrater agreement. We explain why and how the critical values presented must be used differently to answer different research questions. Our intention is to help researchers to interpret interrater agreement under the specified conditions, and, importantly, the results will help researchers to make more appropriate decisions, including those regarding the aggregation of data and more appropriate inferences regarding the interpretation of group phenomena. In what follows, we discuss the AD index, relevant distributions, and interpretive standards for the AD index.

The AD Index of Interrater Agreement

Burke, Finkelstein, and Dusig (1999) introduced the average deviation (AD) as an index of interrater agreement which represents the average absolute deviation in ratings from the mean rating of an item (AD_M)[3], and as such is interpretable in the metric of the original scale. AD_M for an item is calculated as follows:

[3] The average deviation can also be calculated from the median (AD_{Md}) rather than the mean (AD_M). These different versions of the AD index are equal when the mean and median of a distribution are equal, and otherwise they tend to be highly correlated (Burke et al., 1999). Because AD_M is used more often by researchers than AD_{Md}, we focus our

$$AD_{M(J)} = \frac{\sum_{k=1}^{N} \left| x_{jk} - \bar{x}_j \right|}{N}, \quad (1)$$

where N is the number of judges, or observations, of item j, x_{jk} is equal to the *kth* judge's rating of item j, and $\bar{x}_j$ is equal to the mean rating of item j (Burke et al., 1999). The scale $AD_{M(J)}$ is the mean of $AD_{M(j)}$ for essentially parallel items. Because the AD index is a measure of dispersion, lower values indicate greater agreement.

As noted previously, Burke and Dunlap (2002) derived a decision rule for inferring the practical significance of observed AD values. This decision rule has two critical limitations. First, it only addresses assessments of the level of agreement, not the pattern of distributions, which may be theoretically important. With the advance of theories such as DeRue et al.'s (2010) theory of team efficacy dispersion and Harrison and Klein's (2007) theoretical classification of types of diversity, multilevel researchers will need to consider agreement/dispersion as a theoretically meaningful issue. As such, guidelines addressing interpretations of the shapes of distributions are needed. Second, this decision rule applies only when the uniform distribution is the appropriate null distribution. As discussed previously, though the uniform distribution is widely applied, it is thought to be quite often inappropriately applied. There is a mounting push from the scholarly community to justify the choice of a particular null distribution, rather than using the uniform distribution unconditionally, yet too few guidelines exist for researchers who do opt to use alternative null distributions.

Below, we identify the null and theoretical distributions used in our paper. Then we explain how we derived critical values for evaluating the practical significance of interrater

---

paper on AD_M. The Appendix, however, provides the information researchers would need to calculate critical values for AD_{Md} as needed.

agreement in relation to these null distributions. These critical values can be used to assess the level of interrater agreement in regard to data aggregation, which is a within-group assessment, as well as a host of other problems relating to interrater agreement. Further, based on these critical values, we calculated null ranges to be used in relation to studying theoretical problems; that is, assessing the fit between an observed pattern of dispersion and a theoretical distribution.

Interpretive Standards for the AD Index

Here we present our derivations and resulting critical values and null ranges for the AD index given a number of different response distributions. First, the distributions are described in brief. Then, we explain our derivations of interpretive standards for the AD index. Finally, detailed discussions of problems that relate to these theoretical and methodological reasons are presented.

Distributions

The distributions and their methodological and theoretical bases are listed in Table 1. The proportions endorsing each value for 5-point and 7-point scales are listed for each distribution in Tables 2 and 3. Graphical depictions of these distributions are presented in Figures 2-5.

First, we developed critical values for three basic forms of skewed distributions: slight skew, moderate skew, and heavy skew (see Table 2 and Figure 2). Second, while one could model many forms of bimodal distributions, here we simplify our presentation by suggesting two: “moderate” bimodal and “extreme” bimodal. As shown in Table 3 and Figure 3, the size of the subgroups in both cases is 50% of the raters; the difference between the two distributions is that in the moderate bimodal distribution, the subgroups are less divergent than they are in the extreme bimodal distribution. Third, though there are numerous possible ways in which one

could model subgroup distributions, we have simplified our presentation by considering four possibilities based on two dimensions: the size of the subgroup, and the distance between the subgroup ratings and the majority of ratings. These distributions are shown in Table 3, and they are graphically depicted for a 5-point scale in Figure 4. We define a smaller subgroup as 10% of the raters (labeled as "A" in Table 3 and Figure 4) and a more moderately sized subgroup to be 20% of the raters (labeled as "B" in Table 3 and Figure 4). We define extreme distance as the subgroup responses and the majority of responses being on opposite ends of the Likert-type scale, and moderate distance as the subgroup responses being at the midpoint of the scale, while the majority of responses are at one extreme of the scale (these are labeled "extreme" and "moderate" in Table 3 and Figure 4). Finally, we present triangular-shaped, bell-shaped, and uniform distributions in Table 3; they are graphically represented in Figure 5. The triangular-shaped distributions are based on a formula presented by Messick (1982) and the bell-shaped distributions are based on LeBreton and Senter (2008). Note that the upper limits for the uniform distribution (presented in both Tables 2 and 3) are consistent with Burke and Dunlap's (2002) c/6 decision rule for assessing the practical significance of AD, where c is equal to the number of response categories.

Critical Values for Level of Agreement

In order to simplify our derivations, we begin with the basic case of agreement across judges on a single item with respect to two categories.[4] In the case of a dichotomy, AD can be calculated based on the proportion of judges falling into one of the two categories (Burke & Dunlap, 2002)[5]. Based on an upper limit for AD of .35, where .35 or lower represents

[4] We base our work in part on equations presented by Burke and Dunlap (2002).

[5] Burke and Dunlap (2002) demonstrated in their Equation 12 (p. 165) how to calculate AD from a proportion in the case of a dichotomy:

meaningful agreement, Burke and Dunlap (2002) demonstrated that meaningful agreement could be defined as 77% of the judges endorsing one category. Based on a more stringent upper limit of AD, .33, they indicated that meaningful agreement could be defined as 79% agreement.[6] They noted that this notion of 77-79% agreement being meaningful is consistent with many practical examples and problems relating to proportional agreement, such as 60-80% agreement being required for including critical incidents when creating behaviorally anchored rating scales (BARS; Cascio, 1998). Based on Burke and Dunlap's calculations for the AD index, as well as conventional interpretations of meaningful agreement in percent or proportional terms, we adopted a starting value of 80% agreement.

We note that assumptions are necessary to the derivation process. By making ours explicit, readers can readily revise the starting value as needed; yet, we suggest that a starting value of 80% agreement, or 20% disagreement, relative to the AD index is likely to suit most readers' situations. Notably, the value of 20% disagreement is comparable to the upper limits of acceptable disagreement for the AD index for scales that range from 3 to 99 response options (see Burke & Dunlap, 2002); that is, they are comparable in the sense of being approximately

---

$$AD_{(2\ categories)} = 2p(1-p)$$

[6] For the assessment of interrater agreement where judges rate a single target with respect to only two categories (e.g., on a yes-no or agree-disagree dichotomous item format), Burke and Dunlap (2002, p. 164) obtained an upper limit value for AD of .35 using their Equation 9:

$$AD_{UL} = \sqrt{c^2 - {}^{1}/_{24}},$$

where c is equal to the number of categories. When c equals 2, AD_{UL} equals .35. Burke and Dunlap (p. 164) also presented an approximation or simplification of Equation 9, which was their Equation 10:

$$\sqrt{{}^{c^2}/_{25}} = {}^{c}/_{5}.$$

Using an approximation or simplification to Equation 9 of their article (c/5) and dividing this quantity by 1.2, which is the constant by which AD and the standard deviation of responses on an item differ relative to the uniform distribution, yields the value of .33 as the upper limit of AD for a dichotomous item.

equal to the maximum level of allowable disagreement. In other words, our use of the dichotomous case here does not limit the applicability of our derivations to dichotomies.

Given our intent of proposing interrater agreement cut-offs and null ranges for response distributions for Likert-type scales with markedly different dispersion, which have different numbers of response options and reflect a variety of response patterns including non-normal distributions, we next convert a proportion of .80 (or 80%) to a standardized effect size (a correlation coefficient, *r*) to work further with variances as indicators of dispersion. Since non-normal response distributions are expected in many cases for theoretical reasons, we initially employ an arcsine transformation to convert the proportion of .80 to a standardized effect size (i.e., a d-statistic, see Lipsey & Wilson, 2001) and then, use a maximum likelihood transformation of this d-value to obtain a correlation coefficient.

The d-value is computed as the difference between the arcsine of the proportion representing meaningful agreement (i.e., .80) and the arcsine of the proportion representing no agreement (.00) using Lipsey and Wilson's (2001) formula:

$$d = \left(2*\text{arcsine}\left(\sqrt{p_1}\right)\right) - \left(2*\text{arcsine}\left(\sqrt{p_2}\right)\right) \quad (2)$$

The resulting d-value is 2.214. Next, we transform the value of 2.214 to a correlation coefficient via the maximum likelihood formula (Hunter & Schmidt, 2004)[7]:

$$r = \frac{d/2}{\left(1+\left(d/2\right)^2\right)^{1/2}} \quad (3)$$

Given that the proportions in the two groups are unequal (i.e., .80 and .20), the number "2" in Equation 3 is replaced by $1/(p*q)^{1/2}$, where p and q are the proportions in each group. The result

[7] When d is within the range of -.40 to .40, a close approximation of r is d divided by 2. Yet, when d falls outside of this range, the relationship between d and r becomes nonlinear. For the later cases, an accurate approximation of d to r is obtained by the maximum likelihood estimate (see Hunter & Schmidt, 2004, p. 277-279). Because our d of 2.214 is greater than .4, we used Equation 3 to obtain an accurate estimate of r.

is a correlation of approximately .66. By rounding this value up to .7, our derivations continue at the starting point for Burke and Dunlap's (2002) derivations for practical cut-offs for the AD index for the restricted case of the uniform distribution.

We note that arriving at approximately the same value for a correlation as Burke and Dunlap (2002) does not indicate circularity in our derivations, but it does reflect our explicit assumption that the underlying response distribution may meaningfully deviate from normality, thus calling for the arcsine transformation of percent agreement to produce a correlation. As we discuss in the Appendix, assuming that the underlying distribution of responses is normal would call for a probit transformation of the proportion to produce a correlation, and that the resulting value for the correlation would become approximately .8 (Lipsey & Wilson, 2001). Furthermore, Burke and Dunlap's (2002) starting point of .7 for a correlation was, in large part, based on empirical data relating to stability coefficients and correlations based on ratings of targets by alternate sources. Their judgment and ours that a correlation of .7 is a reasonably high correlation is consistent with Cohen (1977) who indicated that correlations greater than or equal to .5 can be considered large.

Next, as is recognized in a number of quantitative fields (e.g., see Burke & Dunlap, 2002; Greene, 1997; Guion, 1998; McCall, 1970; Parsons, 1978), we define a correlation (r) in terms of variances as

$$r = \sqrt{1 - \sigma_e^2 / \sigma_T^2}, \tag{4}$$

where σ_e^2 is the error variance, here representing disagreement, and σ_T^2 is the total variance. Given that the average deviation is a reasonable approximation to the standard deviation (we discuss the more specific relationship below), and the square of the standard deviation is the variance, we can let σ_e^2 equal AD^2. Consistent with James et al.'s (1984, 1993) work, we then set

σ_T^2 to be equal to the variance of the chance responding in the population (σ_{cr-pop}^2). Then, setting *r* equal to .7, as a reasonable value for the correlation in Equation 4, we can rewrite Equation 4 as

$$.7 = \sqrt{1 - AD^2 / \sigma_{cr-pop}^2}. \quad (5)$$

Squaring both sides and solving for the ratio of variances, we obtain

$$AD^2 / \sigma_{cr-pop}^2 = 1 - .7^2 = .51. \quad (6)$$

Rounding .51 to .50 as did Burke and Dunlap (2002) and rewriting Equation 6, we get

$$AD^2 = \sigma_{cr-pop}^2 / 2. \quad (7)$$

We used Equation 7 to calculate AD^2 for the different response distributions we identified. That is, we calculated the variance of each response distribution and then divided the variance by 2 in order to calculate AD^2. By then taking the square root of this resulting value, we calculated *AD*:

$$\sqrt{AD^2} = AD. \quad (8)$$

Recall that in Equation 5, AD^2 was substituted as an approximation for the observed variance; that is to say that *AD* approximates the standard deviation (σ). In fact, the standard and average deviations vary by a constant which is dependent on the specified response distribution. As Burke and Dunlap (2002) noted, for the uniform distribution the σ:*AD* ratio is 1.2. Thus, in order to calculate the upper limits, or critical values, for the AD index, assuming a uniform distribution, they divided *AD* (the result of Equation 8) by 1.2. That is, they corrected for the

difference between σ and AD introduced in Equation 5. The same adjustment is needed here. The resulting value of Equation 8 must be divided by the σ:AD ratio relevant to a given response distribution. Thus, upper limits for acceptable interrater agreement for AD_M must be calculated separately for each null or theoretical response distribution using the following equation:

$$AD_{M_{UL}} = \frac{AD}{(\sigma_e / AD_M)}, \tag{9}$$

where AD is calculated according to Equations 7-8 and AD_M is calculated according to Equation 1.

The resulting critical values are listed in Tables 2 and 3 along with the pattern of responses for each of the distributions identified and the relevant statistics. For use as decision heuristics, we have rounded the critical values in Tables 2 and 3 to two decimal places; they can be applied to individual items or to multi-item scales. For the purpose of assessing the level of agreement, the critical values should be used in the conventional way (Burke & Dunlap, 2002): an observed AD_M value equal to or less than the relevant critical value ($AD_{M_{UL}}$) indicates practically significant agreement. For example, referring to Table 2, under the condition of slight skew for a 5-point scale, $AD_{M_{UL}}$ =.69. Thus, for researchers to infer a practically significant level of observed agreement, one's observed AD_M value must be less than or equal to .69. Based on such an indication of practically significant agreement, researchers would be justified in using the mean score as an indicator of the group's standing on a construct of interest and as a data point for further, multilevel analysis.

Null Ranges for Pattern of Dispersion

In addition to developing critical values, in response to recent advances in multilevel theory, we also developed null ranges to facilitate researchers' ability to assess how well the shape of an observed distribution fits a theoretically specified distribution. This issue of comparing the pattern of observed dispersion with a theoretical distribution is analogous to judging the goodness-of-fit between one's data and the theoretical response distribution. Cortina and Folger (1998) described tests of goodness-of-fit as a matter of accepting the null hypothesis of no statistically significant difference between observed data and theoretical models. Here, we are dealing with practical significance rather than statistical significance meaning that goodness-of-fit in this context is a matter of concluding that there is no *meaningful* difference between an observed distribution and the theoretically specified response distribution.

The values for AD_M shown in Tables 2 and 3 quantify the dispersion of different distributions. Therefore, if an observed value is equal to the relevant tabled AD_M value, then the pattern of observed dispersion should fit perfectly with the pattern of theoretical dispersion. It is unlikely, however, that observed and tabled values will perfectly match; thus, the question becomes what is the "null range"? In other words, how far can an observed value be from the tabled value before researchers must conclude that their observed distribution has a poor fit with the theoretical distribution?

Analogous to Greenwald's (1975) discussion of how to accept a null hypothesis gracefully (also see discussions by Cashen & Geiger, 2004; Cortina & Folger, 1998), researchers would need to decide in advance of collecting data what magnitude of effect, in this case, the magnitude of AD_M, would be considered nontrivial. We suggest defining this magnitude as the difference between the expected AD_M value for a distribution and the respective upper limit for that distribution. While the decision to specify this magnitude is arguably somewhat arbitrary, it

is nevertheless made in advance of collecting data and tied to our derivations for assessments of practical agreement. Consistent with Greenwald's (1975) arguments about establishing a null range for the formal test of a null hypothesis, this minimum magnitude of AD_M that the researcher is willing to consider nontrivial is then the boundary of the null range. That is, for observed AD_M values, this magnitude would be the difference between the tabled AD_M value and the relevant tabled critical value, $AD_{M_{UL}}$. The general equation for establishing a null range for a theoretical response distribution is as follows:

$$AD_{M_{nullrange}} = AD_M \pm (AD_M - AD_{M_{UL}}) / w, \quad (10)$$

where w is used to define the width of the null range. Herein, following Greenwald's (1975, pp. 16-18) logic regarding establishing a "two-tailed" null range that is symmetric around the zero point of a test statistic, we define w as equal to 2:

$$AD_{M_{nullrange}} = AD_M \pm (AD_M - AD_{M_{UL}}) / 2. \quad (11)$$

The resulting values are presented in Tables 2 and 3. For use as decision heuristics, we have rounded the lower (symbolized by a minus sign) and upper (symbolized by a plus sign) limits of the null range to two decimal points.

Although we present null ranges that are symmetrical around the expected AD_M value, researchers can readily define w and the width of the null range relative to the purposes of their investigations. In these cases, larger values for w will result in smaller, more conservative null ranges than those reported in Tables 2 and 3 for the respective response distributions. In addition, researchers may desire to consider the construction of non-symmetrical, one-tailed null ranges for some types of theoretical response distributions. As with the use of critical values, the researcher may desire to consider *a priori* several theoretical response distributions when making

judgments about whether the observed and theoretical response distributions are meaningfully different.

Using the null ranges to gauge the goodness-of-fit between an observed and a theoretical distribution is straightforward. Using the previous example of a slightly skewed null distribution and a 5-point scale, an observed AD_M of .98 would suggest a perfect match between the observed and theoretical distributions (see Table 2). Yet, how can a researcher interpret an observed AD_M of .70? The relevant range is .84 to 1.12.[8] Thus, a researcher who observes an AD_M of .70 would conclude that the observed distribution is meaningfully different from the theoretical distribution. That is, the observed AD_M of .70 falls outside of the null range.

What researchers would do after determining a lack of fit would depend upon the theoretical context. In some cases it may be that a lack of fit suggests that the phenomena researchers are attempting to study are not represented in the data. This eventuality is analogous to researchers who do multilevel research finding a lack of agreement such that aggregation to a higher level of analysis cannot be justified (e.g., Chan, 1998). Or, it may be the case that shapes of observed distributions are compared to multiple theoretically specified distributions. While .7 does not fall into the null range for slight skew, it does fall into the range for moderate skew. In this case, the researcher would be able to categorize the group as a "moderate skew" group and make theoretically-based predictions accordingly. More broadly, researchers can use the null ranges provided in Tables 2 and 3 in order to classify groups according to the pattern of their distributions of scores, and then based on this classification, make theoretically derived predictions about group outcomes. These null ranges and those relevant to the other distributions discussed below can be used relative to a single item or a scale.

[8] Note that if readers were to plug the tabled values into Equation 9, they would arrive at a slightly different range due to rounding error.

It is important to note that researchers must visually check the observed distribution of responses. For instance, the direction of skew may be of theoretical relevance. Because the AD index is calculated via absolute values, the direction of skew cannot be determined from AD values. Quantifying agreement as well as visually checking the direction of skew is necessary. This point holds for other distributions discussed as well.

Distribution Choice

Below, we discuss examples of when these different response distributions might be relevant (see also Table 1). Note that we do not assume that only one distribution is relevant in any given research context; rather, as others have suggested (e.g., James et al., 1984), we think it is reasonable that multiple distributions may be appropriate. We organize our discussion by first considering the issue of level of agreement and then considering the issue of pattern of dispersion. Within these sections, we refer to distributions defined by skew (Table 2) and those defined by kurtosis and variance (Table 3).

Level of Agreement

A number of response biases suggest that the appropriate null distribution is a skewed distribution. James et al. (1984) and LeBreton and Senter (2008) have discussed the likelihood of leniency and social desirability in contexts of assessing interrater agreement. Leniency may apply, for instance, in the performance appraisal domain where subordinates tend to judge their supervisors in relatively positive terms (Schriesheim, 1981). Klein, Conn, Smith, and Sorra (2001) found social desirability to be applicable in a survey of organizational members' workplace perceptions. Agreement among members was related to the social desirability of the survey items (e.g., "The supervisor to whom I report praises me for excellent performance" and "My work here is enjoyable"; Klein et al., p. 11). To the extent that these biases are expected to

be strong versus weak, and to the extent that multiple biases are expected to be relevant, researchers could utilize moderately to heavily skewed distributions as their null distribution.

Though skewed distributions have most often been suggested as alternatives to the uniform null distribution, other distributions are relevant as well. Likert-type response formats that convey or have different informational value may result in subgroups or small to moderate percentages of respondents using particular response options. For instance, Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991) showed participants responded differently to the question "how successful would you say you have been in life?" when the 11-point scale ranged from -5 to 5 rather than 0 to 10, even though the anchors were identical (not at all successful to extremely successful). For the former scale, 34% endorsed -5 to 0; for the latter scale, 13% endorsed 0 to 5. Schwarz (1999) argued that the question of success is somewhat ambiguous in that success could be marked by the presence of positive features or the absence of negative features, and that participants use the scale numbers as well as the anchors to interpret the items. In addition, Lindell and Brandt (1997) discussed the possibility of distinct factions among raters due to characteristics such as clinical orientations in assessments of psychotherapy, raters' academic disciplines in rating research proposals, raters' functional department in ratings of organizational climate, and so on. The possibility of such factions might call for the use of a bimodal response distribution as the baseline distribution for assessing level of agreement among a set or raters. We present four different subgroup distributions and two different bimodal distributions (see Table 3). In addition, triangular-shaped or bell-shaped distributions[9] are applicable null distributions if one expects raters to succumb to the central tendency bias (e.g., James et al., 1984; LeBreton & Senter, 2008). For instance, James et al. (1984, p. 91) suggested

[9] Here we are dealing with multinomial distributions, and, as a result, we refer to them as triangular-shaped and bell-shaped as they are respectively similar in a figurative sense to the normal probability distribution.

that the central tendency bias may occur "…when judges are purposefully cautious or evasive because responses to items are not collected on a confidential basis, and political reasons exist for not departing from the neutral alternatives on the scales." They also suggested that naïve and unmotivated participants may exhibit the central tendency bias when responding to ambiguous or complicated items.

Finally, while the uniform distribution has been described as an often inappropriate null distribution, there are circumstances under which it is the appropriate null distribution. It is applicable if no rater bias is expected. It may also be applicable if raters face conceptual ambiguity. For instance, Heidemeier and Moser (2009) found that raters demonstrated less agreement in job performance ratings when the work being evaluated was less straightforward; that is, there was less agreement regarding white-collar work and work high in job complexity compared to agreement regarding blue-collar work and work low in job complexity.

Pattern of Dispersion

There are also theoretical bases for modeling agreement on most of these distributions. DeRue et al. (2010) provided an example theoretical basis for choosing a bimodal distribution as a theoretical response distribution: equally sized subgroups within teams that judge team efficacy differently will have mixed effects on team effectiveness by impairing social processes, but enhancing task processes. They went on to propose that the greater the divergence between the subgroups, the more negative the effect on team effectiveness will be. In discussing maximum separation diversity, such as diversity in team members' judgments of team efficacy, Harrison and Klein (2007) discussed an extreme bimodal distribution, where subgroups exist on opposite ends of a continuum. Consistent with DeRue et al. (2010), they argued that this extreme bimodal

distribution would have negative outcomes: reduced cohesiveness, interpersonal conflict, distrust, and decreased task performance.

Related to bimodal distributions are unimodal distributions that have distinct subgroups. From a theoretical perspective, DeRue et al. (2010) discussed "minority belief" dispersion where one team member rates team efficacy differently than the other team members. We previously reported their proposition that when minority belief dispersion is characterized by one individual rating team efficacy lower than everyone else, the effect on team effectiveness will be negative. DeRue et al. (2010) also theorized about the opposite distribution: one individual rates team efficacy more highly than the other team members. They proposed that this pattern of dispersion, which is the mirror-image of the first scenario, will have mixed effects on team effectiveness because the dispersion is likely to impair social processes, but enhance task processes.

Finally, DeRue et al. (2010) provided an example of when the uniform distribution would be theoretically specified as the expected response distribution: fragmentation, characterized by a uniform distribution of team efficacy beliefs, should augment team effectiveness by positively impacting social and task processes. Their argument is based on the idea that fragmented teams may communicate more effectively than other teams because they do not have subgroups, coalitions, and factions that can hinder effective communication in teams, and they are motivated to create a shared understanding of team efficacy. These teams are likely to openly discuss issues like goals and expectations that can help in teams' task-related processes, as well as helping to establish a shared belief about team efficacy. Harrison and Klein (2007) proposed similarly positive effects for variety diversity, such as diversity in educational background, when

it is at a maximum level, which is characterized by a uniform distribution: more creativity, greater innovation, higher decision quality, more task conflict, and increased unit flexibility.

Conclusion

Given numerous calls for researchers who use interrater agreement indices to stop their unconditional use of the uniform response distribution, a primary purpose of our study was to provide researchers with guidelines for using alternative null distributions and theoretical distributions to make judgments about practical significance, when addressing both methodological and theoretical issues. In doing so, we derived critical values for a variety of response distributions that vary in terms of skew, kurtosis, and variance. We also discussed how to use the critical values shown in Tables 2 and 3 differently depending on whether one seeks to ascertain the *level* of agreement or the *pattern* of dispersion. While the question of the level of agreement is familiar, the question of the pattern of dispersion is more novel, but likely to become more and more important with advances in multilevel theory and research. The current paper stands to promote such conceptual advances.

Although we focused the substantive discussion of interrater agreement problems on data aggregation in relation to the *level* of agreement and team efficacy dispersion and diversity in relation to the *pattern* of dispersion, the derived critical values and null ranges can be applied to numerous other research questions. For instance, the alternative null distributions and critical values could assist in addressing interrater agreement questions related to job analysts' ratings of task items for a job, or judges' ratings of critical or cut-off scores on the items of a test (e.g., using the Angoff method whereby cut-off scores are based on subject matter experts' estimates of the probability that a competent person will respond to an item accurately; e.g., see Hudson & Campion, 1994) just to name a few types of pertinent research questions. As another example,

the notion of a theoretical distribution could be used to specify the demographic makeup of a community (e.g., racial/ethnic composition in terms of percentages within each category), thereby permitting the quantification of demographic similarity/dissimilarity between employees and residents (i.e., the difference between an observed AD_M value and the relevant tabled value for a theoretical distribution). Quantifying the effects of community demographic similarity in this manner may meaningfully extend the measurement and study of employee-community racial/ethnic similarity from the individual level of analysis (e.g., see Avery, McKay, & Wilson, 2008; Brief, Umphress, Dietz, Burrows, Butz, & Scholten, 2005) to the organization or business unit levels of analysis. Importantly, irrespective of the group phenomena under study, the AD index itself and the derived null ranges provide a means for tracking or studying expected changes in group phenomena possibly relative to stages of group development or shocks that the group might encounter. Further, practical significance critical values could be similarly developed for other interrater agreement indices.

Future research should also address the problem of assessing the statistical significance of AD values relative to a variety of null or theoretical distributions. As discussed earlier, the work to date in this area is limited. Burke and Dunlap (2002) and Dunlap et al. (2003) used an approximate randomization test to establish statistical significance cut-offs for AD for judges' ratings of a single item relative to the uniform distribution. Cohen et al. (2009) built upon this work to establish statistical significance cut-offs for AD for judges' agreement on multi-item scales relative to the uniform distribution and a slightly skewed distribution. In order to assist researchers in inferring whether levels of agreement (i.e., AD values) are due to chance, cut-off values for statistical significance should be established relative to more distributions, such as

those identified in Tables 2 and 3. Without additional guidelines, researchers are likely to continue to over-rely on the uniform distribution when making inferences about their data.

In closing, we emphasize that the practical guidelines presented herein are just that: guidelines. As others have advised, it is important that researchers take a common sense approach to interpreting observed agreement. Speaking in terms of whether interrater agreement is sufficient to justify the aggregation of individual level data, LeBreton and Senter (2008, p. 836) asserted that "…the value used to justify aggregation ultimately should be based on a researcher's consideration of (a) the quality of the measures, (b) the seriousness of the consequences resulting from the use of aggregate scores, and (c) the particular composition model to be tested." James et al. (1984), in addressing the problem of uncertainty over which null distribution applies in a given situation, suggested interpreting observed agreement on the basis of several null distributions: "The rationale here is that even though we cannot pinpoint a particular null with a high degree of confidence, we can place bounds on the most likely types of nulls and thereby increase the likelihood that the true null lies somewhere in this range of distributions" (p. 95).

Similarly, we urge researchers to consider their particular circumstances when assessing interrater agreement and to consider the use of a range of critical values based on several different null or theoretical distributions. For instance, researchers should consider whether they have missing data (e.g., see Newman & Sin, 2009). Our guidelines do not account for systematically missing data and, thus, may be sensitive to this problem, particularly in the cases of certain distributions, such as the bimodal distribution, which may appear as a unimodal distribution if data are systematically missing from one of the two subgroups. In other cases, researchers may need to apply a null or theoretical distribution not included in Tables 2 and 3, or

they may need to adjust the starting value for interrater agreement of 80% agreement used in the present derivations. Moving away from 80% agreement or considering another transformation of percent agreement for derivational purposes, such as a probit transformation of a proportion, will result in more stringent or more lenient critical values and decisions concerning interrater agreement depending on whether one adjusts this value upward or downward, or whether one employs a more versus less conservative transformation of the proportion, such as arcsine versus probit transformations. Recognizing the possibility that the research context may dictate the consideration of other assumptions or response distributions than those used in the study, we present in the Appendix a general procedure for researchers to use in establishing critical values and null ranges based on other assumptions not considered herein. In this regard, our proposed guidelines offer a uniform and parsimonious means for studying interrater agreement given a variety of methodological and theoretical problems.

References

Avery, D. R., McKay, P. F., & Wilson, D. C. (2008). What are the odds? How demographic similarity affects the prevalence of perceived employment discrimination. *Journal of Applied Psychology, 93*, 235-249.

Bledow, R. & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62,* 229–258.

Borucki, C. C., & Burke, M. J. (1999). An examination of service-related antecedents to retail store performance. *Journal of Organizational Behavior, 20,* 943-962.

Brief, A. P., Umphress, E. E., Dietz, J., Burrows, J. W., Butz, R. M., & Scholten, L. (2005). Community matters: Realistic group conflict theory and the impact of diversity. *Academy of Management Journal, 48*, 830-844.

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

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5,* 159-172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2,* 49-68.

Cascio, W. F. (1998). *Applied psychology in human resource management*. Upper Saddle River, NJ: Prentice Hall.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7*(2), 151-167.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234-246.

Cohen, A., Doveh, E., & Nahum-Shani, I. (2009). Testing agreement for multi-item scales with the indices $r_{WG(J)}$ and $AD_{M(J)}$. *Organizational Research Methods, 12,* 148-164.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods, 1,* 334-350.

Dawson, J. F., Gonzalez-Roma, V., Davis, A., & West, M. A. (2008). Organizational climate and climate strength in UK hospitals. *European Journal of Work and Organizational Psychology, 17,* 89-111.

DeRue, D. S., Hollenbeck, J. R., Ilgen, D. R., & Feltz, D. (2010). Efficacy dispersion in teams: Moving beyond agreement and aggregation. *Personnel Psychology, 63,* 1-40.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{WG} and average deviation indexes. *Journal of Applied Psychology, 88,* 356-362.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13,* 668-689.

Grant, A. M., & Mayer, D. M. (2009). Good soldiers and good actors: Prosocial and impression management motives as interactive predictors of affiliate citizenship behaviors. *Journal of Applied Psychology, 94*, 900-912.

Greene, W. H. (1997). *Econometric analysis*. Upper Saddle River, NJ: Prentice-Hall.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1-20.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum.

Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review, 32,* 1199-1228.

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology, 94,* 353-370.

Hudson, J. P., & Campion, J. E. (1994). Hindsight bias in an application of the Angoff method for setting cutoff scores. *Journal of Applied Psychology, 79,* 860-865.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

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

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

Katz-Navon, T., Naveh, E., & Stern, Z. (2009). Active learning: When is more better? The case of resident physicians' medical errors. *Journal of Applied Psychology, 94*, 1200-1209.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology, 86,* 3-16.

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19,* 195-229.

Kline, T. J. B., & Hambley, L. A. (2007). Four multi-item interrater agreement options: Comparisons and outcomes. *Psychological Reports, 101,* 1001-1010.

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

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3-90). San Francisco: Jossey-Bass.

Kreiner, G. E., Hollensbe, E. C., & Sheep, M. L. (2009). Balancing borders and bridges: Negotiating the work-home interface via boundary work tactics. *Academy of Management Journal, 52*, 704-730.

Lawrence, K. L., Lenk, P., & Quinn, R. E. (2009). Behavioral complexity in leadership: The psychometric properties of a new instrument to measure behavioral repertoire. *Leadership Quarterly, 20*, 87-102.

Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement, 21*, 271-278.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815-852.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Lüdtke, O., & Robitzsch, A. (2009). Assessing within-group agreement: A critical examination of a random-group resampling approach. *Organizational Research Methods, 12,* 461-487.

McCall, R. B. (1970). *Fundamental statistics for psychology*. New York: Harcourt, Brace, & World, Inc.

Messick, D. M. (1982). Some cheap tricks for making inferences about distribution shapes from variances. *Educational and Psychological Measurement, 42,* 749-758.

Meyer, R. D., Mumford, T. V., & Campion, M. A. (2010, August). *The practical consequences of null distribution choice on rwg*. Paper presented at the annual meeting of the Academy of Management, Montreal, Canada.

Newman, D. A., & 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 nonresponses. *Organizational Research Methods, 12,* 113-147.

Nicklin, J. M., & Roch, S. G. (2009). Letters of recommendation: Controversy and consensus from expert perspectives. *International Journal of Selection and Assessment, 17*, 76-91.

Parsons, R. (1978). *Statistical analysis*. New York: Harper & Row.

Roberson, Q. M., Sturman, M. C., & Simons, T. L. (2007). Does the measure of dispersion matter in multilevel research? A comparison of the relative performance of dispersion indexes. *Organizational Research Methods, 10,* 564-588.

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior, 7,* 1-37.

Schriesheim, C. A. (1981). The effect of grouping or randomizing items on leniency response bias. *Educational and Psychological Measurement, 41,* 401-411.

Schwarz, N. (1999). Self-reports: How questions shape the answers. *American Psychologist, 54*, 93-105.

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55,* 570-582.

Smith-Crowe, K., & Burke, M.J. (2003). Interpreting the statistical significance of observed AD interrater agreement values: Corrections to Burke and Dunlap (2002). Organizational Research Methods, 6, 129-131.

Takeuchi, R., Chen, G., & Lepak, D. P. (2009). Through the looking glass of a social system: Cross-level effects of high performance work systems on employees' attitudes. *Personnel Psychology, 62*, 1-29.

Trougakos, J. P., Beal, D. J., Green, S. G., & Weiss, H. M. (2008). Making the break count: An episodic examination of recovery activities, emotional experiences, and positive affective displays. *Academy of Management Journal, 51,* 131-146.

Van Kleef, G. A., Homan, A. C., Beersma, B., Van Knippenberg, D., Knippenberg, B. V., & Damen, F. (2009). Searing sentiment or cold calculation? The effects of leader emotional displays on team performance depend on follower epistemic motivation. *Academy of Management Journal, 52*, 562-580.

Wagner, S. M., Rau, C., & Lindermann, E. (2010). Multiple informant methodology: A critical review and recommendations. *Sociological Methods and Research, 38,* 582-618.

Walumbwa, F. O., & Schaubroeck, J. (2009). Leader personality traits and employee voice behavior: Mediating roles of ethical leadership and work group psychological safety. *Journal of Applied Psychology, 94,* 1275-1286.

Table 1

Example Theoretical and Methodological Bases for Different Response Distributions

Distribution	Theoretical Basis	Methodological Basis
Skew	• Social Interaction • Work Interdependence • Shared Schemas • Maximum Disparity Diversity	• Social Desirability • Leniency
Bimodal	• Equal Subgroups • Maximum Separation Diversity	• Factions
Subgroup	• Minority Belief, or Unequal Subgroups	• Response Formats and Unintended Question Interpretation
Triangular/ Bell-Shaped		• Central Tendency
Uniform	• Fragmentation • Maximum Variety Diversity	• Absence of Bias • Conceptual Ambiguity

Table 2

Critical Values and Null Ranges for AD_M Given Distributions Defined by Skew

Distribution	Proportion Endorsing Each Value							σ	AD_M	$\frac{\sigma}{AD_M}$	Critical Values $AD_{M_{UL}}$	Null Ranges AD_M -	Null Ranges AD_M +
	1	2	3	4	5	6	7						
5-Point Scale													
Slight Skew	.05	.15	.20	.35	.25			1.34	0.98	1.18	0.69	0.84	1.12
Moderate Skew	.00	.10	.15	.40	.35			0.90	0.70	1.36	0.49	0.60	0.80
Heavy Skew	.00	.00	.10	.40	.50			0.44	0.60	1.11	0.42	0.51	0.69
Uniform	.20	.20	.20	.20	.20			2.00	1.20	1.18	0.85	1.02	1.38
7-Point Scale													
Slight Skew	.05	.08	.12	.15	.20	.25	.15	2.92	1.44	1.19	1.02	1.23	1.65
Moderate Skew	.00	.06	.10	.14	.28	.22	.20	2.09	1.16	1.25	0.82	0.99	1.33
Heavy Skew	.00	.00	.05	.10	.15	.30	.40	1.39	0.94	1.25	0.66	0.80	1.08
Uniform[a]	.14	.14	.14	.14	.14	.14	.14	4.00	1.71	1.17	1.21	1.46	1.97

Note. The critical values were calculated without restricting decimal places, but they were rounded to two decimal places for reporting purposes. The only exception was AD_M, which was restricted to two decimal places when inputted into the calculations.
[a]The proportions are rounded such that they do not sum to 1. For this scale, equal proportions summing to 1 require 15 decimal places.

Table 3

Critical Values and Null Ranges for AD_M Given Distributions Defined by Kurtosis and Variance

Distribution	Proportion Endorsing Each Value							σ^2	AD_M	$\frac{\sigma}{AD_M}$	Critical Values $AD_{M_{UL}}$	Null Ranges AD_M	
	1	2	3	4	5	6	7					-	+
5-Point Scale													
Moderate Bimodal[a]	.00	.50	.00	.50	.00			1.00	1.00	1.00	0.71	0.85	1.15
Extreme Bimodal[a]	.50	.00	.00	.00	.50			4.00	2.00	1.00	1.41	1.71	2.29
Moderate Subgroup A[ab]	.00	.00	.10	.00	.90			0.36	0.36	1.67	0.25	0.31	0.41
Extreme Subgroup A[ab]	.10	.00	.00	.00	.90			1.44	0.72	1.67	0.51	0.61	0.83
Moderate Subgroup B[ab]	.00	.00	.20	.00	.80			0.64	0.64	1.25	0.45	0.55	0.73
Extreme Subgroup B[ab]	.20	.00	.00	.00	.80			2.56	1.28	1.25	0.91	1.09	1.47
Triangular-Shaped	.11	.22	.34	.22	.11			1.32	0.88	1.31	0.62	0.75	1.01
Bell-Shaped	.07	.24	.38	.24	.07			1.04	0.76	1.34	0.54	0.65	0.87
Uniform	.20	.20	.20	.20	.20			2.00	1.20	1.18	0.85	1.02	1.38
7-Point Scale													
Moderate Bimodal[a]	.00	.50	.00	.00	.00	.50	.00	4.00	2.00	1.00	1.41	1.71	2.29
Extreme Bimodal[a]	.50	.00	.00	.00	.00	.00	.50	9.00	3.00	1.00	2.12	2.56	3.44
Moderate Subgroup A[ab]	.00	.00	.00	.10	.00	.00	.90	0.81	0.54	1.67	0.38	0.46	0.62
Extreme Subgroup A[ab]	.10	.00	.00	.00	.00	.00	.90	3.24	1.08	1.67	0.76	0.92	1.24
Moderate Subgroup B[ab]	.00	.00	.00	.20	.00	.00	.80	1.44	0.96	1.25	0.68	0.82	1.10
Extreme Subgroup B[ab]	.20	.00	.00	.00	.00	.00	.80	5.76	1.92	1.25	1.36	1.64	2.20
Triangular-Shaped	.06	.13	.19	.24	.19	.13	.06	2.50	1.26	1.25	0.89	1.08	1.44
Bell-Shaped	.02	.08	.20	.40	.20	.08	.02	1.40	0.84	1.41	0.59	0.72	0.96
Uniform[c]	.14	.14	.14	.14	.14	.14	.14	4.00	1.71	1.17	1.21	1.46	1.97

Note. The critical values were calculated without restricting decimal places, but they were rounded to two decimal places for reporting purposes. The only exception was AD_M, which was restricted to two decimal places when inputted into the calculations. [a]"Moderate" and "extreme" refer to the distance between subgroups. [b]"A" and "B" refer to the differential proportion of subgroup responses. [c]The proportions are rounded such that they do not sum to 1. For this scale, equal proportions summing to 1 require 15 decimal places.

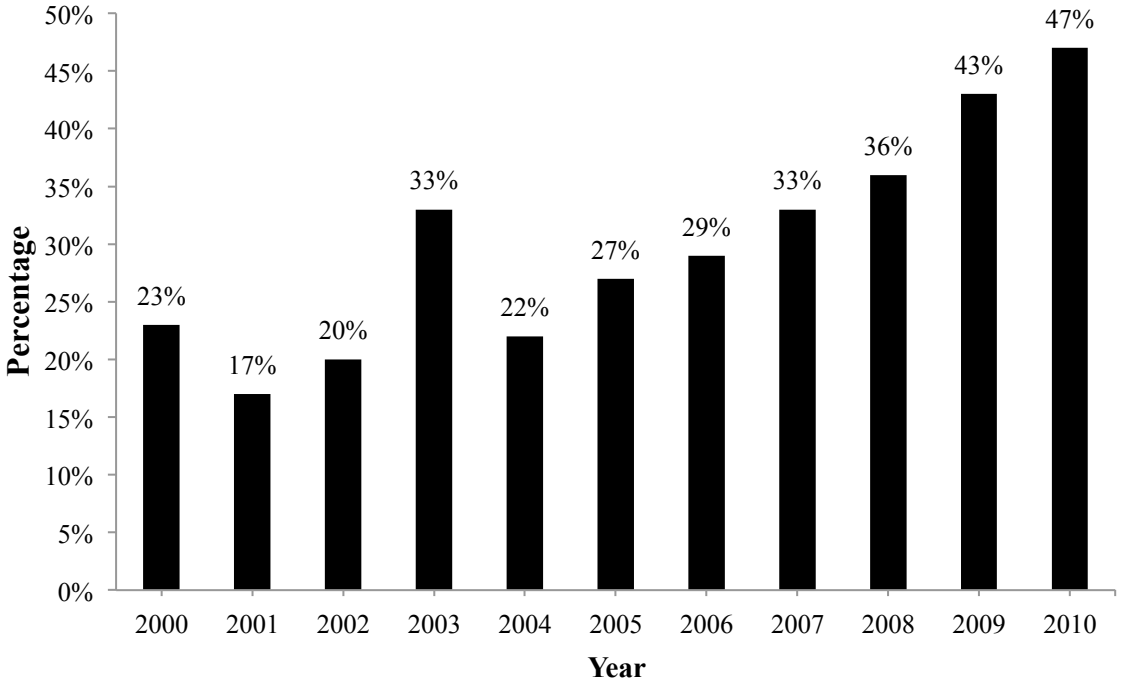

Figure 1. Percentage of articles published in *Personnel Psychology* and the *Journal of Applied Psychology* that used interrater agreement statistics, including r_{WG}, AD, ICC, percent agreement, and Cohen's kappa.

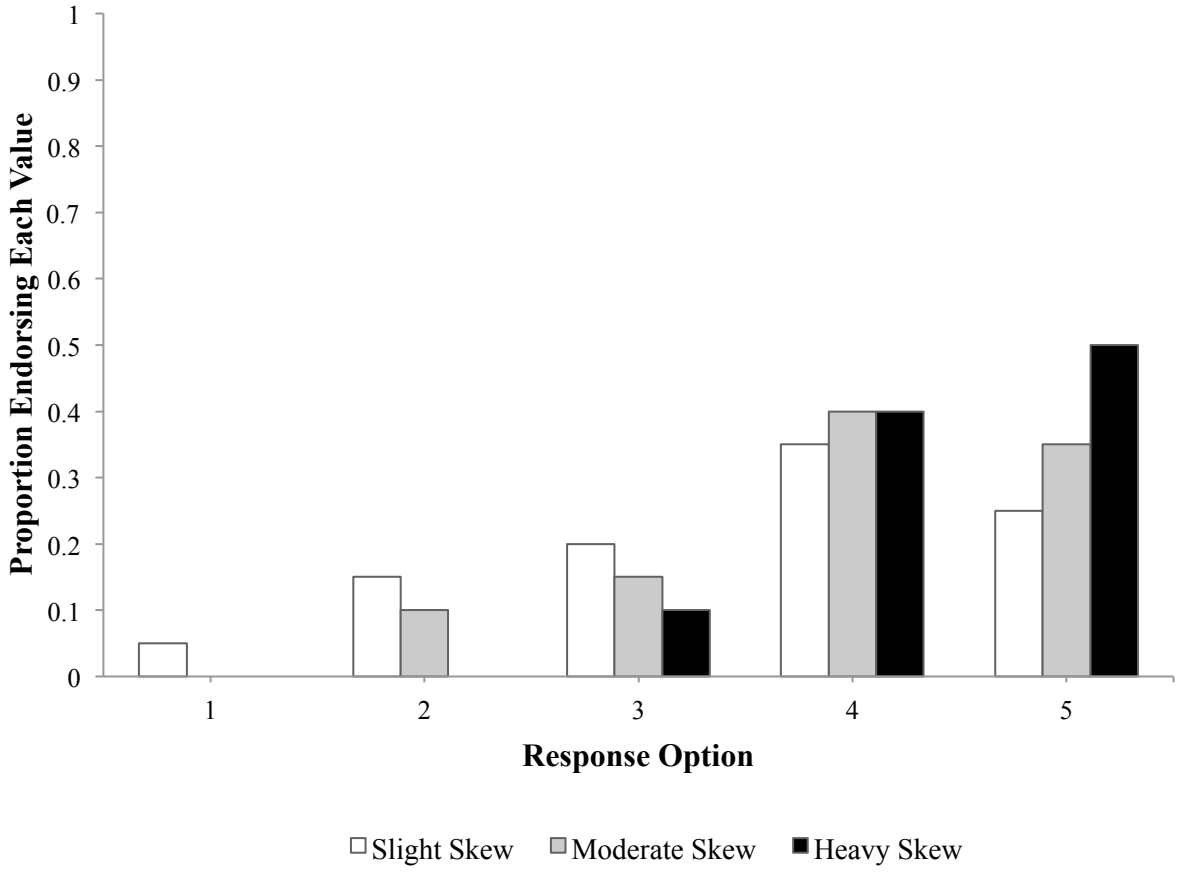

Figure 2. Slight, moderate, and heavy skew distributions for a 5-point scale.

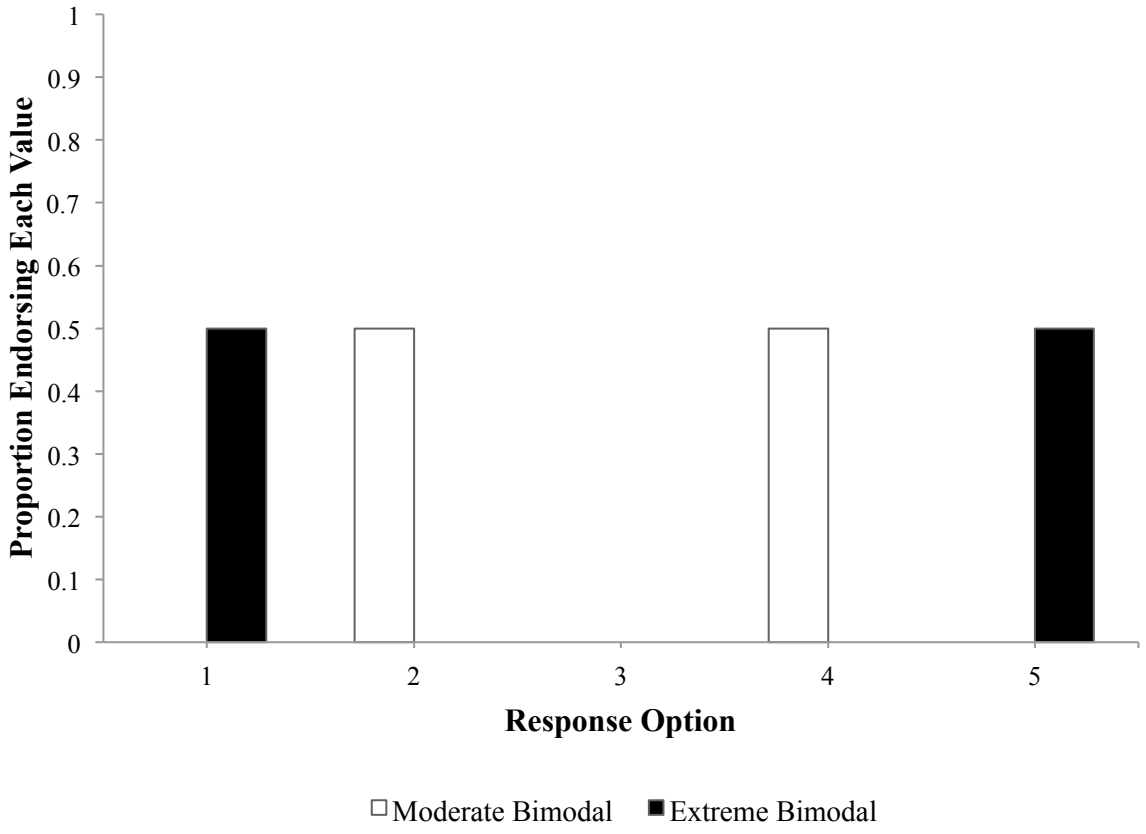

Figure 3. Bimodal distributions for a 5-point scale.

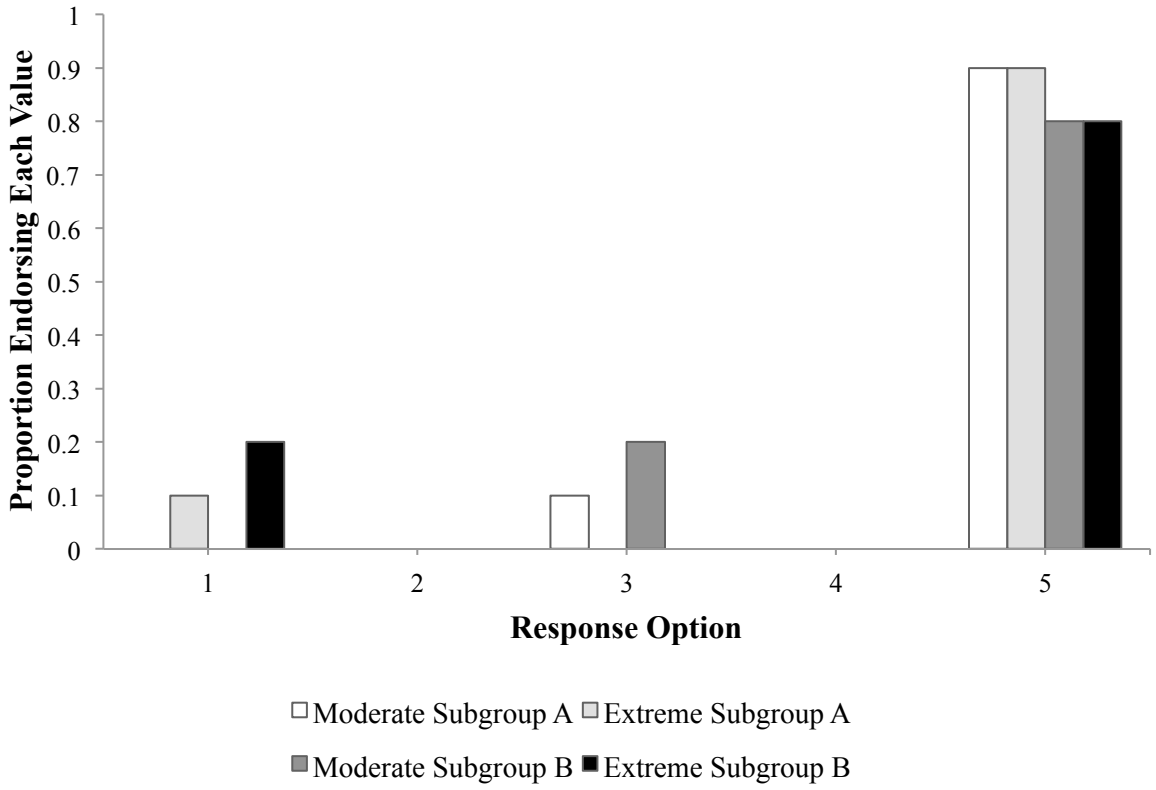

Figure 4. Subgroup distributions for a 5-point scale.

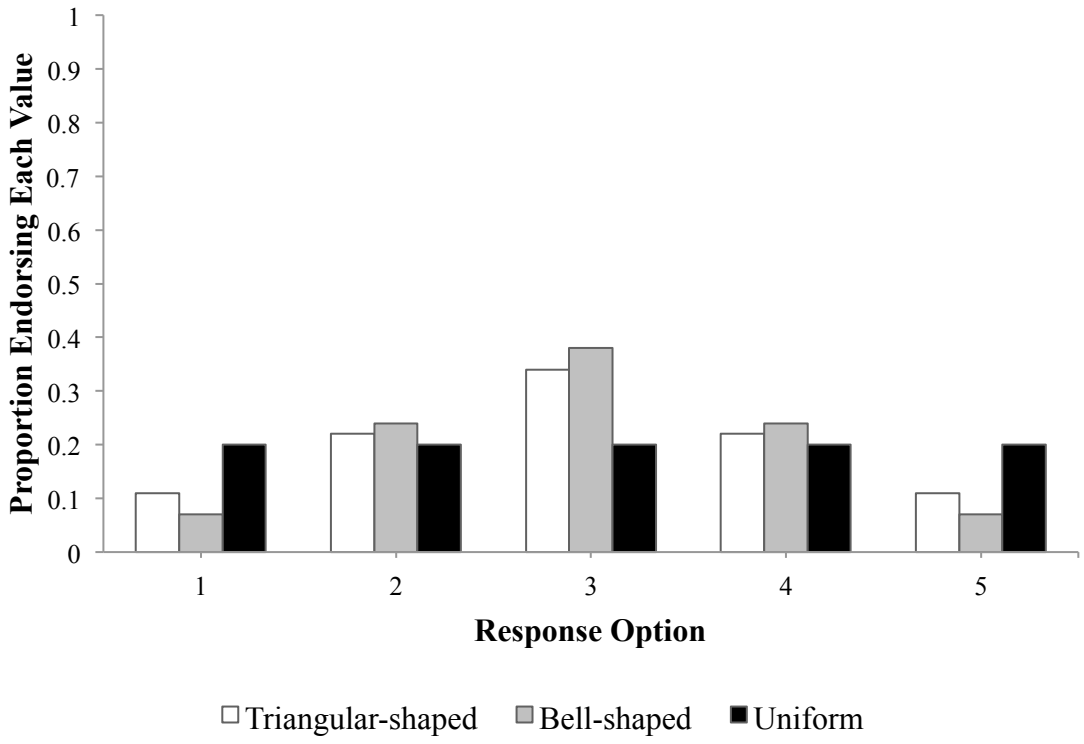

Figure 5. Triangular-shaped, bell-shaped, and uniform distributions for a 5-point scale.

Appendix

Calculating Critical Values and Null Ranges for the AD Index

Although we presented a number of different response distributions in Tables 2 and 3, researchers may find that their methodologically or theoretically specified response distribution is not listed. In this case, researchers can follow our procedures to calculate the relevant critical values and null ranges. The first step is to express the response distribution in terms of the proportion of individuals endorsing each value of a scale as we did in Tables 2 and 3. The second step is to calculate the variance for the specified distribution. Third, as per Equation 7, divide the variance by 2 to calculate AD^2; then, take the square root of the resulting value (i.e., *AD*, see Equation 8). Finally, as per Equation 9, divide *AD* by the σ:AD_M ratio to calculate $AD_{M_{UL}}$; follow Equation 10 to calculate the null range (see also Equation 11). The agree.exe program available at http://www.tulane.edu/~dunlap/psylib.html can be used to calculate AD_M for items or multi-item scales. The calculations conducted by the software are based on Burke and Dunlap (2002) and Dunlap et al. (2003). Note that the "actual variance" reported by the software is calculated for a sample; rather, our calculations are based on the variance calculated for a population.

In addition, researchers may find that starting with a percent agreement of 80% and using a correlation of .7 (within Equation 5) is either too lenient or too stringent given their particular circumstances. In such cases, researchers can derive their own critical values and null ranges based on different initial assumptions. Beginning with either a different percent agreement or continuing derivations with another value for the correlation will produce different cut-offs and null ranges. One can substitute another reasonable value in Equation 5 and follow the sequence through Equation 10 to arrive at new critical values and null ranges (see also Equation 11). For

example, replacing .7 with .8 in Equations 5 and 6, results in $1-.8^2 = .36$. Thus, Equation 7 would be rewritten such that the variance would be divided by 2.78. To calculate their own critical values then, researchers using an $r = .8$ would calculate the variance of a given distribution and divide that variance by 2.78 (as per the revised version of Equation 7). Then, they would follow Equations 8-9 to calculate $AD_{M_{UL}}$, and Equation 10 to calculate the null range (see also Equation 11).

There are several reasons for which researchers may want to use cut-offs associated with a correlation of .8. For instance, from our starting point of defining meaningful agreement as 80% agreement, one could use a probit transformation to convert this proportion to an effect size. A probit transformation of the proportion may be called for if the underlying distribution of scores is expected to be normally distributed. Also the probit transformation may be a particularly good choice in estimating a standardized effect from proportions if the cut-point between the two groups is in the tail portion of a skewed distribution (Lipsey & Wilson, 2001). A probit transformation of a proportion of .8 will produce a correlation equal to .78501 (Lipsey & Wilson), which rounds to .8. Also, as discussed in Burke and Dunlap (2002), a correlation of .8 would correspond to a high level of stability in scores. Substituting .8 for .7 in Equation 6 and then following the remainder of the equations, one would arrive at more stringent critical values than those presented in Tables 2 and 3. We present this more stringent set of criteria in Tables A1 and A2.

Finally, some researchers may want to use AD calculated from the median (AD_{Md}) rather than the mean (AD_M). These different versions of the AD index are equal when the mean and median of a distribution are equal, and otherwise they tend to be highly correlated (Burke et al., 1999). Though AD_M has been used more often by researchers, Burke et al. (1999) argued that

AD_{Md} is more sensitive in detecting agreement since the median of a distribution is the point at which the sum of the absolute deviations are the most minimal, and more minimal deviations indicate higher agreement. AD_{Md} for an item is calculated as follows:

$$AD_{Md(j)} = \frac{\sum_{k=1}^{N} \left| x_{jk} - Md_j \right|}{N}, \tag{12}$$

where Md_j is equal to the median rating of item j and all other notations are consistent with those in Equation 1. The scale $AD_{Md(J)}$ is the mean of $AD_{Md(j)}$ for essentially parallel items. The upper limit for AD_{Md} would be calculated as follows:

$$AD_{Md_{UL}} = \frac{AD}{(\sigma_e / AD_{Md})}, \tag{13}$$

where AD is calculated according to Equations 5-8. Finally, the null range for AD_{Md} would be calculated as follows:

$$AD_{Md_{nullrange}} = AD_{Md} \pm (AD_{Md} - AD_{Md_{UL}}) / w, \tag{14}$$

where w is used to define the width of the null range. For the same reasons given previously in the discussion of Equation 11, we suggest defining w as equal to 2.

Table A1

Critical Values and Null Ranges for AD_M Given Distributions Defined by Skew and r=.8

Distribution	Proportion Endorsing Each Value							σ	AD_M	$\frac{\sigma}{AD_M}$	Critical Values $AD_{M_{UL}}$	Null Ranges AD_M	
	1	2	3	4	5	6	7					-	+
				5-Point Scale									
Slight Skew	.05	.15	.20	.35	.25			1.34	0.98	1.18	0.59	0.78	1.18
Moderate Skew	.00	.10	.15	.40	.35			0.90	0.70	1.36	0.42	0.56	0.84
Heavy Skew	.00	.00	.10	.40	.50			0.44	0.60	1.11	0.36	0.48	0.72
Uniform	.20	.20	.20	.20	.20			2.00	1.20	1.18	0.72	0.96	1.44
				7-Point Scale									
Slight Skew	.05	.08	.12	.15	.20	.25	.15	2.92	1.44	1.19	0.86	1.15	1.72
Moderate Skew	.00	.06	.10	.14	.28	.22	.20	2.09	1.16	1.25	0.69	0.92	1.39
Heavy Skew	.00	.00	.05	.10	.15	.30	.40	1.39	0.94	1.25	0.56	0.75	1.13
Uniform[a]	.14	.14	.14	.14	.14	.14	.14	4.00	1.71	1.17	1.03	1.37	2.06

Note. The critical values were calculated without restricting decimal places, but they were rounded to two decimal places for reporting purposes. The only exception was AD_M, which was restricted to two decimal places when inputted into the calculations.
[a]The proportions are rounded such that they do not sum to 1. For this scale, equal proportions summing to 1 require 15 decimal places.

Table A2

Critical Values and Null Ranges for AD_M Given Distributions Defined by Kurtosis and Variance and r=.8

Distribution	Proportion Endorsing Each Value							σ	AD_M	$\frac{\sigma}{AD_M}$	Critical Values $AD_{M_{UL}}$	Null Ranges AD_M	
	1	2	3	4	5	6	7					-	+
5-Point Scale													
Moderate Bimodal[a]	.00	.50	.00	.50	.00			1.00	1.00	1.00	0.60	0.80	1.20
Extreme Bimodal[a]	.50	.00	.00	.00	.50			4.00	2.00	1.00	1.20	1.60	2.40
Moderate Subgroup A[ab]	.00	.00	.10	.00	.90			0.36	0.36	1.67	0.22	0.29	0.43
Extreme Subgroup A[ab]	.10	.00	.00	.00	.90			1.44	0.72	1.67	0.43	0.58	0.86
Moderate Subgroup B[ab]	.00	.00	.20	.00	.80			0.64	0.64	1.25	0.38	0.51	0.77
Extreme Subgroup B[ab]	.20	.00	.00	.00	.80			2.56	1.28	1.25	0.77	1.02	1.54
Triangular-shaped	.11	.22	.34	.22	.11			1.32	0.88	1.31	0.53	0.70	1.06
Bell-shaped	.07	.24	.38	.24	.07			1.04	0.76	1.34	0.46	0.61	0.91
Uniform	.20	.20	.20	.20	.20			2.00	1.20	1.18	0.72	0.96	1.44
7-Point Scale													
Moderate Bimodal[a]	.00	.50	.00	.00	.00	.50	.00	4.00	2.00	1.00	1.20	1.60	2.40
Extreme Bimodal[a]	.50	.00	.00	.00	.00	.00	.50	9.00	3.00	1.00	1.80	2.40	3.60
Moderate Subgroup A[ab]	.00	.00	.00	.10	.00	.00	.90	0.81	0.54	1.67	0.32	0.43	0.65
Extreme Subgroup A[ab]	.10	.00	.00	.00	.00	.00	.90	3.24	1.08	1.67	0.65	0.86	1.30
Moderate Subgroup B[ab]	.00	.00	.00	.20	.00	.00	.80	1.44	0.96	1.25	0.58	0.77	1.15
Extreme Subgroup B[ab]	.20	.00	.00	.00	.00	.00	.80	5.76	1.92	1.25	1.15	1.54	2.30
Triangular-shaped	.06	.13	.19	.24	.19	.13	.06	2.50	1.26	1.25	0.76	1.01	1.51
Bell-shaped	.02	.08	.20	.40	.20	.08	.02	1.40	0.84	1.41	0.50	0.67	1.01
Uniform[c]	.14	.14	.14	.14	.14	.14	.14	4.00	1.71	1.17	1.03	1.37	2.06

Note. The critical values were calculated without restricting decimal places, but they were rounded to two decimal places for reporting purposes. The only exception was AD_M, which was restricted to two decimal places when inputted into the calculations. [a][c]"Moderate" and "extreme" refer to the distance between subgroups. [b]"A" and "B" refer to the differential proportion of subgroup responses. [c]The proportions are rounded such that they do not sum to 1. For this scale, equal proportions summing to 1 require 15 decimal places.