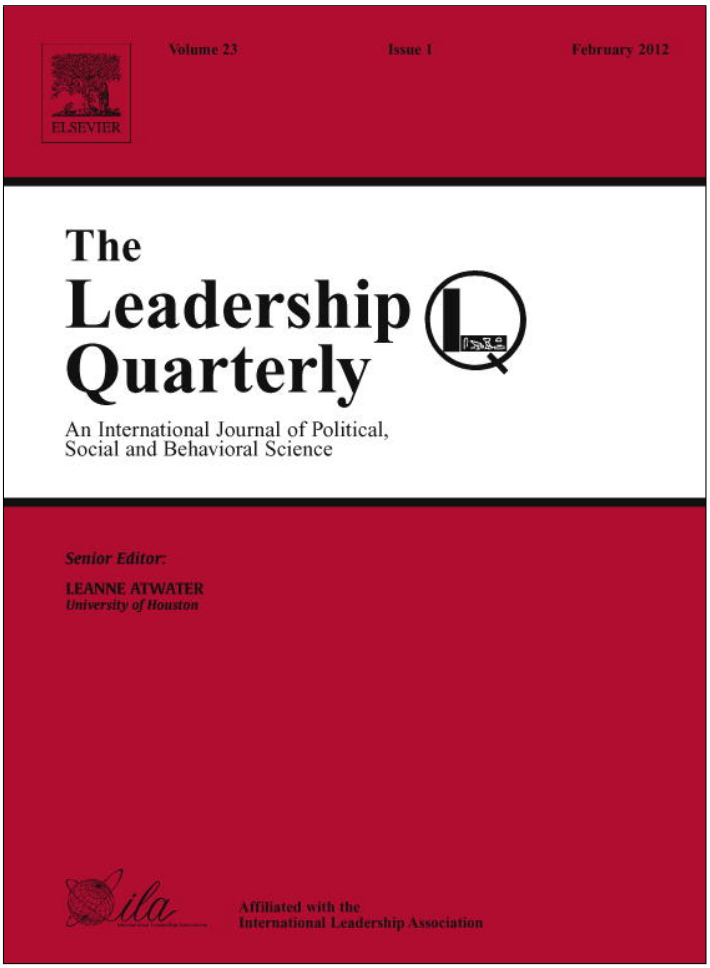

The Leadership Quarterly 23 (2012) 66–80

Contents lists available at SciVerse ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Within-group agreement: On the use (and misuse) of r_{WG} and $r_{WG(J)}$ in leadership research and some best practice guidelines ☆

Torsten Biemann [a,*], Michael S. Cole [b], Sven Voelpel [c,d]

[a] University of Cologne, Germany, Personnel Economics and Human Resource Management, 50672 Cologne, Germany
[b] Texas Christian University, USA
[c] Jacobs University Bremen, Germany
[d] EBS Business School, Germany

ARTICLE INFO

Available online 3 December 2011

Keywords:
Data aggregation
Multilevel methods
Within-group agreement

ABSTRACT

Multilevel leadership researchers have predominantly applied either *direct consensus* or *referent-shift consensus* composition models when aggregating individual-level data to a higher level of analysis. Consensus composition assumes there is sufficient within-group agreement with respect to the leadership construct of interest; in the absence of agreement, the aggregate leadership construct is untenable. At the same time, guidelines to help leadership researchers make decisions regarding data aggregation issues have received little explicit attention. In particular, a discussion of how data aggregation decisions can enhance or obscure a study's theoretical contribution – a central focus of this article – has not been addressed thoroughly. Recognizing that empirical generalization depends on the accuracy with which aggregation decisions are applied, we revisit the often neglected assumptions associated with the most common agreement statistic used to justify data aggregation — r_{WG} and $r_{WG(J)}$ (James, Demaree, and Wolf, 1984). Thereafter, using a dataset published as part of a *Leadership Quarterly* special issue (Bliese, Halverson, & Schriesheim, 2002), we highlight the potential misuse of r_{WG} and $r_{WG(J)}$ as the sole statistic to justify aggregation to a higher level of analysis. We conclude with prescriptive implications for promoting consistency in the way multilevel leadership research is conducted and reported.

1. Introduction

The inclusion of multiple levels of analysis in the study of leadership phenomena has gained increasing importance (Yammarino & Dansereau, 2008; Yammarino, Dionne, Chun, & Dansereau, 2005). The majority of this research relies on survey data gathered from individuals and then aggregated to the leader or group-level of analysis. In such models, an aggregate-level leadership measure is created by averaging subordinates' assessments of their leaders' behavior. For example, in a study conducted by Rubin, Munz, and Bommer (2005), these researchers examined leaders' emotion recognition ability and how this ability influenced their leadership behavior. Whereas Rubin et al. (2005) focused on specific leaders (i.e., leader level of analysis), others have utilized higher level entities in their leadership research. Bass, Avolio, Jung, and Berson (2003) examined how leader behaviors directed at unit members as a whole (commonly known as leadership climate) were related to unit performance (i.e., unit-level of analysis). Finally, Dickson, Resick, and Hanges (2006) studied the linkages between organizational-level factors and effective leadership (i.e., organizational-level of analysis). Although each of these studies examined a leadership phenomenon at a different level of analysis, they all utilized a consensus-based composition model (Chan, 1998) to operationalize their unit-level leadership

☆ We thank Joyce Bono, Stephanie Castro, Jeremy Dawson, Janaki Gooty, Astrid Homan, and Frank Walter for their helpful comments.
* Corresponding author. Tel.: +49 221 470 7955.
E-mail address: biemann@wiso.uni-koeln.de (T. Biemann).

1048-9843/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.leaqua.2011.11.006

construct.[1] Another unifying factor in these three studies is their reliance on James et al.'s (1984) r_{WG} procedure to justify aggregating individual leadership perceptions to the proposed level of analysis. As we discuss later in this paper, the level of agreement or homogeneity across individual group members' judgments is a central consideration for consensus composition constructs (Chan, 1998).

This approach of first justifying aggregation vis-à-vis interrater agreement (i.e., r_{WG} index for measures with a single item; $r_{WG(J)}$ for multiple item measures) empirically and then testing the hypothesized relationships between the higher level constructs and criteria is a common and accepted practice.[2] The application of r_{WG}-based indices has, however, come under scrutiny by research methodologists. In brief, r_{WG} was initially introduced as an index of interrater reliability (i.e., the relative consistency in ratings provided by multiple judges of multiple targets). Criticism by Schmidt and Hunter (1989) resulted in a relabeling of r_{WG} as an index of interrater agreement (Kozlowski & Hattrup, 1992), as it does not conform to the concept of reliability in standard measurement theory (James, Demaree, & Wolf, 1993). Further, scholars have criticized the widely-applied cut-off criterion of $r_{WG} = .70$ as purely arbitrary (Castro, 2002; Charnes & Schriesheim, 1995; Cohen, Doveh, & Eick, 2001; Dunlap, Burke, & Smith-Crowe, 2003; Lance, Butts, & Michels, 2006; LeBreton, James, & Lindell, 2005). Moreover, when computing r_{WG} and $r_{WG(J)}$ values, the observed within-group variances are compared to an expected variance under the null hypothesis of no agreement. Nevertheless, scholars have argued that there is no clear-cut definition of a response corresponding to no agreement (Cohen, Doveh, & Nahum-Shani, 2009). Consequently, the ambiguity in choosing the most appropriate null response pattern (i.e., distribution) is often noted as a major limitation of r_{WG} and $r_{WG(J)}$.

The ideas in this paper are not intended to contribute to the ongoing methodological discussion of r_{WG} and $r_{WG(J)}$ but, rather, to help leadership researchers think through their data aggregation decisions in a more explicit and systematic way. To this end, we sought to integrate relevant measurement, design, and analytical considerations to provide a nontechnical tutorial and methodological resource when attempting to justify aggregation of lower level leadership data to a higher level of analysis. It is particularly significant that there is no readily available integrative framework for promoting consistency in the way in which this research is conducted and reported. Consequently, there is considerable variation in leadership researchers' justification of data aggregation. This has implications, since leadership studies employing different procedures for summarizing individual-level data in order to operationalize consensus composition constructs may yield contradictory and noncomparable findings.

By offering some clarifying heuristics to help leadership researchers develop and hone their data-aggregation decision-making abilities, we hope to address the above problems and contribute to multilevel leadership research in several important ways. We start by revisiting the underlying assumptions and applications of r_{WG} and $r_{WG(J)}$ and, explain why the *typical* multilevel leadership study using r_{WG} indices (to justify data aggregation) is subject to potential criticism. We then underscore the potential misuse of r_{WG} and $r_{WG(J)}$ agreement indices with an empirical example. Finally, we provide best practice guidelines for researchers interested in exploring aggregate leadership phenomena and journal referees commissioned to review such work. The brief list of practices we identify, along with a discussion of important caveats, allows for a more fine-grained discussion of the data aggregation options than was previously possible.

2. Composition models and within-group agreement

An ever-increasing number of leadership researchers have applied multilevel frameworks in their work and, thus, proper measurement of leadership phenomena that emerge from lower levels is a perennial concern (e.g., Dansereau & Yammarino, 2006; Yammarino & Dansereau, 2008). To guide such efforts, researchers have used Chan's (1998) typology of composition models to specify the functional relationship between phenomena at different levels of analysis. Essentially, Chan's typology provides researchers with a framework for mapping the transformation of constructs across analysis levels.

In aggregating lower level scores to index a higher level leadership construct, a majority of leadership researchers have adopted either a *direct consensus* or a *referent-shift consensus* model (Cole & Bedeian, 2007). *Direct consensus* models use averaged individual members' responses to operationalize group-level scores (Chan, 1998). For example, followers may be asked to rate their leader's charisma (e.g., "I am inspired by my leader's vision for the group"). Conversely, *referent-shift consensus* models require individual group members to respond to survey items in reference to a higher level unit (Chan, 1998). Researchers might ask followers of a leader to rate the degree to which they agree with the statement "My group is inspired by our leader's vision for the group". Thus, rather than asking followers about their individual perceptions, referent-consensus models incorporate a different referent (i.e., a group as a whole). Nonetheless, both forms of consensus composition require group members to be homogenous regarding the target construct (e.g., leader charisma). That is, "in the absence of substantial within-unit agreement" the unit-level measure comprised of individuals' aggregated responses "has no construct validity" (Klein, Conn, Smith, & Sorra, 2001, p. 4). To determine whether there is sufficient agreement among followers' responses to represent a group's (e.g., subordinates reporting to a single leader or members of a team) standing on a given leadership variable, interrater agreement indices such as the r_{WG} and $r_{WG(J)}$ are computed and compared to threshold values.

[1] Although there are potential differences between work groups and teams (Chan, 1998, p. 235), for simplicity we view them similarly as a clustering of individuals who are interdependent, share a set of common expectations or hierarchical structuring, and who interact with one another as if in a group. As such, we use the terms "group," "unit," and "team" interchangeably.

[2] The application of a consensus composition model is typically a two step process. Step 1 involves computing an agreement index in order to demonstrate that averaging individual members' responses yields a valid unit-level construct. Step 2 (assuming high agreement is observed) involves testing hypothesized relationships using multilevel statistical packages such as Hierarchical Linear and Nonlinear Modeling (HLM), R, or SAS. Given that our objective is to underscore the notion of how data aggregation decisions can enhance or obscure a study's theoretical contribution, we focus on issues pertaining to Step 1.

3. r_{WG} and $r_{WG(J)}$

Given that interrater agreement refers to the absolute consensus in the scores that respondents provide (Cohen et al., 2009; James et al., 1993; Kozlowski & Hattrup, 1992; Tinsley & Weiss, 1975), when espousing a consensus composition model, it is important for leadership researchers to compute (some form of) r_{WG} to "address whether scores furnished by judges are interchangeable or equivalent in terms of their absolute value" (LeBreton & Senter, 2008, p. 816). Building on this basis, we take the position that it is inappropriate to condemn r_{WG} and $r_{WG(J)}$ statistics when it is the researcher who is not applying the statistics properly. Indeed, as will be discussed, few researchers regularly implement James et al.'s (1984) advise on computing r_{WG} and $r_{WG(J)}$ statistics. Not only researchers new to the leadership field, but also accomplished leadership researchers who are new to multilevel leadership research should benefit from an explicit discussion of r_{WG}-based statistics.

The application of r_{WG}-based indices is based on the belief that each target (e.g., manager) has a true score on the assessed construct (e.g., transformational leadership). Any variance among raters (e.g., followers) is assumed to be error variance. Accordingly, interrater agreement can be estimated by comparing the observed variance to the variance expected when there is complete lack of agreement among raters (i.e., random responding). The decision on whether to calculate r_{WG} or $r_{WG(J)}$ is solely determined by the measurement instrument employed. Whereas r_{WG} is the within-group agreement for a single item, $r_{WG(J)}$ combines the r_{WG} estimates for each item of a multi-item measure. According to James et al. (1984), r_{WG} is calculated as:

$$r_{WG} = 1 - \frac{S_{jk}^2}{\sigma_{EU}^2}. \tag{1}$$

The r_{WG} index relates the within-group variance of a single item j in a group of k raters (S_{jk}^2) to an expected variance that assumes all ratings were due to random responding (σ_{EU}^2). This index has also been extended to measures that essentially comprise parallel items. James et al. (1984) suggest the multi-item $r_{WG(J)}$, which is calculated as:

$$r_{WG(J)} = \frac{J * \left(1 - \frac{S_{.k}^2}{\sigma_{EU}^2}\right)}{1 + (J-1) * \left(1 - \frac{S_{.k}^2}{\sigma_{EU}^2}\right)}. \tag{2}$$

The $r_{WG(J)}$ index applies the Spearman–Brown prophecy formula to include the number of items in the calculation of within-group agreement. Thus, J is the number of items in a measure and $S_{.k}^2$ the average variance of the J items in a group of k raters. For example, if a measure is comprised of five items – each exhibiting an r_{WG} of 0.5 – working through Eq. (2) yields an $r_{WG(5)} = 5*0.5/(1+4*0.5) = 0.83$. The rationale for applying a correction that increases with the number of items is that measurement error is generally reduced by the inclusion of additional items (e.g., Lord, Novick, & Birnbaum, 1968). It follows that $r_{WG(J)}$ is typically larger in magnitude than is the mean of J r_{WG} indices of the same construct (James et al., 1984). Consequently, the probability of exceeding a widely-applied cut-off criterion of .70 (cf. Lance et al., 2006) will usually increase with the number of items.

4. Assumptions and the null distribution of no agreement

Various assumptions are made for both single item (r_{WG}) and multi-item ($r_{WG(J)}$) forms of interrater agreement. First, the measures being employed must have "acceptable psychometric properties" (James et al., 1984, p. 85). This includes construct validity and reliability. Second, a measure's response options should approximate equal-interval measurement (James et al.). Third, because these agreement indices were intended to be used when analyzing data with discrete response formats, James et al. recommend using a 5- or 7-point response ramp. The use of fewer response options (e.g., a 3-point response format) can result in artificially low estimates of interrater agreement. Additionally, the agreement index for multi-item measures should only be applied to measures with "essentially parallel indicators of the same construct" (James et al., p. 88). The multi-item measure is thus assumed to tap a unidimensional construct.

James et al. (1984) initially developed r_{WG} and $r_{WG(J)}$ to provide accurate and interpretable estimates of rater agreement that also allow for the controlling of response biases (e.g., central tendency and social desirability). Others have likewise cautioned against the universal positivity bias that occurs in attribution-making (e.g., Mezulis, Abramson, Hyde, & Hankin, 2004). As noted by James et al., the controlling of response biases can be accomplished by the *careful selection* of the underlying assumption associated with the null distribution of no agreement (i.e., the expected variance [σ_{EU}^2] estimation). They recognized, however, that it is virtually impossible to determine if a specific observed distribution is the result of true scores or due to some form of response bias. Consequently, they suggested that one should identify a small but inclusive set of null distributions and use them to compute a range of agreement scores in which true agreement is most likely to occur. It should be noted that leadership researchers (mirroring the general management literature) have generally ignored the recommendations by James et al. In so doing, they have opted to represent the null distribution of no agreement with the rectangular (uniform) distribution. A key problem with the rectangular (uniform) null distribution is that it assumes all answering options have the same probability of being selected (Cohen et al., 2009; LeBreton & Senter, 2008). Fig. 1a depicts a rectangular (uniform) distribution. As shown, if a 5-point response continuum is used, each response option has an equal chance (i.e., 20%) of being selected by a rater; hence, the distribution is flat and rectangular.

5. Implications for multilevel leadership research

According to James et al. (1984), there are many instances when random responding (i.e., no agreement) will not correspond to a rectangular distribution. If a common tendency among raters is to select a socially desirable response option rather than one reflective of their true beliefs, the subsequent appearance of high interrater agreement is indicative of a response bias and is not reflective of true agreement among raters (James et al., 1984, 1993). In a study of leadership, this situation may occur, for example, when subordinates (a) exhibit a positive leniency in describing their managers' transformational leadership behavior or (b) select a neutral response option because they wish to evade a particular question set (e.g., survey items designed to assess abusive supervision). In both instances, the clustering of subordinates' ratings in the observed distribution might be (incorrectly) assumed to reflect true agreement, although this clustering of ratings is due to respondents' rating biases. Consequently, selecting the rectangular (uniform) null distribution to derive an expected variance is inappropriate and could result in spuriously high estimates of within-group agreement. In the positive leniency example, a skewed distribution should be used to estimate the expected variance, whereas in the evasive condition, a triangular or central tendency distribution is the more appropriate null distribution (see James et al.). Fig. 1b and 1c respectively depicts a moderately skewed and normal distribution.

Thus, it has been shown that blindly using the rectangular (uniform) null distribution may obscure the true distribution of members' responses (Brown & Hauenstein, 2005; Lebreton, Burgess, Kaiser, Atchley, & James, 2003). Echoing James et al.'s (1984) recommendation, methodologists have continued to emphasize that failing to consider alternative null distributions may cast doubt on a study's findings insofar as the groups' agreement scores based on the rectangular (uniform) null distribution will often yield inflated estimates (e.g., LeBreton & Senter, 2008). With the above-cited research in mind, a potential criticism of contemporary leadership research is that, despite growing recognition that the rectangular (uniform) null distribution may not be applicable in many situations, the vast majority of studies still rely on it (for exceptions, see Bono & Judge, 2003; Bono, Foldes, Vinson, & Muros, 2007; Liao & Chuang, 2007; Schriesheim, Cogliser, & Neider, 1995; Shamir, Zakay, Breinin, & Popper, 1998; Walker, Smither, & Waldman, 2008).

6. An empirical example of $r_{WG(J)}$

For illustrative purposes, we present an empirical example based on a set of data used in the 2002 special issue of *The Leadership Quarterly* on "multilevel issues in leadership" (Bliese et al., 2002). Our aim is not only to provide a brief illustration in the context of leadership, but to also show the potential consequences of failing to consider alternative null distributions when within-group agreement statistics, such as the $r_{WG(J)}$, needs to be computed. As recommended by James et al. (1984), we compare $r_{WG(J)}$ values derived solely from a rectangular (uniform) distribution with a range of values based on a small set of alternative distributions. If sizeable differences are observed in this example, one can infer potential implications may exist for other leadership streams (e.g., LMX or transformational leadership) that routinely use r_{WG} and $r_{WG(J)}$ statistics.

In our opinion, the following example is a fairly common situation that leadership researchers face. The sample consists of 2042 soldiers nested in 49 groups (i.e., U.S. Army Companies). To begin, let us assume that, based on theory and research purposes, the direct consensus composition model (Chan, 1998) is an appropriate approach to operationalize *leadership climate*, *task significance*, and *group hostility*. Let us also assume that we have good reason to posit the hypothesis: leadership climate (i.e., average leadership perceptions in an Army Company) will moderate the relationship between the absolute levels (i.e., aggregated mean value) of group task significance and group hostility (see Bliese et al., 2002). Accordingly, individual members ($n = 2042$) of the 49 Army Companies were asked to assess their leaders' consideration and support on an 11-item measure, using a 5-point response ramp. Task significance was assessed using a three-item measure, while hostility was assessed using a five-item measure; responses were on a 5-point response continuum. Since each construct is assessed using a multi-item measure, the ensuing analyses focus on $r_{WG(J)}$.

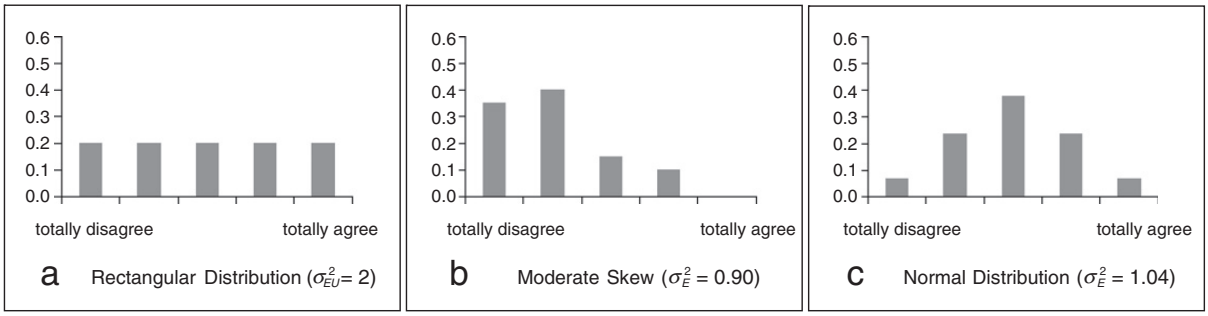

Fig. 1. Theoretical null distributions associated with a 5-point Likert-type response scale.

7. Aggregation results of army company data

7.1. The leadership climate example

In choosing to operationalize leadership climate using a direct consensus model, within-group agreement is a prerequisite for aggregating soldiers' ratings to the Company level. In other words, before we move to hypothesis testing, we must demonstrate that members of each Army Company are homogenous with respect to their leadership ratings. As an initial step, we choose a null distribution and compute an expected variance $\sigma^2_{EU} = (A^2 - 1)/12$. The subscript EU reflects the expected error "E" variance based on a uniform "U" distribution, whereas "A" corresponds to the number of response options. Thus, utilizing a rectangular (uniform) distribution which assumes that all answering options have the same probability, we obtain an expected variance (σ^2_{EU}) of 2 (i.e., $[5^2 - 1]/12 = 2$). Second, to obtain r_{WG} indices for each item of the measure (see Eq. 1), we compare the observed within-group variance from the 49 groups to the expected variance and subtract the obtained value from 1. Finally, the mean $r_{WG(J)}$ is calculated according to the average r_{WG} values and the number of items on the measure (see Eq. 2). Results indicate there is high agreement in terms of leadership climate, with a mean $r_{WG(11)}$ of 0.87 (ranging from 0.77 to 0.94). Based on the widely-applied cut-point of .70 (cf. Lance et al., 2006), it would seem that one could justify aggregating these data to the Company level of analysis.

In contrast, a more detailed inspection of the item-level variances reveals a different picture. That is, by computing the total variance for each item of a measure, it is possible to compare a mean within-group variance based on actual sample data with a mean within-group variance derived from randomly assigned groups. The 11 leadership items' total item variances range from 1.06 to 1.73, with an average of 1.33. Consider the leadership item with a total variance of 1.06. If we were to draw randomly from the total sample of Army Companies and calculate the item's variance (based on this random draw of individuals), the value of the *expected* variance for this randomly created pseudo group is 1.06. That is, an item's total variance is equivalent to the observed variance that we can expect for random groups comprised of members from different Army Companies. This is important to note, because the groups' observed variance is incorporated into the numerator of both the r_{WG} and the $r_{WG(J)}$. With this in mind, now consider another group of randomly selected individuals with an expected average variance of 1.33 (as obtained in the sample data). In working through Eq. (2) for leadership climate, the expected $r_{WG(J)}$ of a random group is, $11*(1-1.33/2)/(1+10*(1-1.33/2)) = 0.85$. In other words, when we randomly assign soldiers to pseudo groups, we still obtain an average $r_{WG(J)}$ that clearly exceeds the (arbitrarily set) threshold of 0.70.

Remember that the actual sample data yielded a mean $r_{WG(J)}$ of 0.87, and that the pseudo group based on random allocations yielded only a slightly lower mean $r_{WG(J)}$ of 0.85. It follows that the agreement in the actual Army Companies is higher than the agreement obtained in the pseudo groups, thereby demonstrating "true" within-group agreement. Nevertheless, the absolute difference is relatively small and, thus, one could question whether these Army Companies should have their individual-level data aggregated to the specified unit of analysis.

7.2. The task significance and group hostility examples

We followed the same procedures in calculating the $r_{WG(J)}$ for each multi-item measure. The mean $r_{WG(3)}$ for task significance was 0.58 in the actual (Army Company) groups compared to an expected mean $r_{WG(3)}$ of 0.56 for the pseudo groups. The mean $r_{WG(5)}$ for group hostility was 0.56 in the actual groups compared to an expected mean $r_{WG(5)}$ of 0.54 for the pseudo groups. Similar to the leadership climate measure, the $r_{WG(J)}$ values corresponding to task significance and hostility are higher in the actual groups than one would expect from random assignment into pseudo groups. It is therefore appropriate to conclude that clustering effects exist for both variables because the *actual* $r_{WG(J)}$ values are higher than one would expect for pseudo groups (see Bliese & Halverson, 1996; 2002). Note that despite comparable clustering effects across all three variables (namely, leadership climate, task significance, and hostility) as expressed by the differences between the actual and pseudo groups' $r_{WG(J)}$ values, the often used 0.70 threshold does not justify aggregating the task significance and hostility data to represent variables at the Army Company level of analysis.

7.3. On the use of alternative null distributions

Recall that r_{WG}-based estimates derived from the rectangular (uniform) null distribution are likely to yield inflated values (LeBreton & Senter, 2008). One should thus view the r_{WG} values reported above with caution, since they are based on this rectangular distribution. We therefore decided to identify a more inclusive set of null distributions and use them to compute a range of agreement scores in which true agreement is most likely (James et al., 1984). Table 1 provides $r_{WG(J)}$ values for leadership climate, task significance, and group hostility. The $r_{WG(J)}$ values derived from a rectangular or uniform distribution (described earlier) should be viewed as an upper limit; the $r_{WG(J)}$ values based on the alternative null distributions (termed "measure-specific" in Table 1) can be interpreted as a theoretical lower bound of within-group agreement.

In short, we identified two possible alternative null distributions for leadership climate. One might argue that a slightly negative skew in the null distribution is most likely, due in part to a leniency bias of the followers providing the ratings. Alternatively, one could also argue that a normal null distribution is more likely given that leaders' characteristics in question (i.e., positive and supportive behaviors) are non-rectangular. In other words, we might expect the "true" distribution of leaders' behaviors to assume a shape that more closely follows a normal distribution (alternatively known as the bell curve) because of a "true" bell curve in the population. For example, not all supervisors make exceptional leaders; it is more natural for supervisors' leadership

scores to be clustered around the overall mean because, in the population of all possible supervisors, the majority of supervisors will be "average" leaders. As shown in Table 1, the lower bound r_{WG} estimate for leadership climate when using a slight skew and normal distribution is 0.41 and 0.10, respectively. Hence, the average within-group agreement score for leadership climate is most likely between 0.10 and 0.87.

We identified one alternative null distribution for task significance to set the lower bound estimate for within-group agreement. Given that the study respondents were US soldiers, we anticipated that many respondents would view their tasks as important (i.e., a slight negative skew). Similarly, we also anticipated that the majority of respondents would be reluctant to report instances of within-group hostility (i.e., a moderate positive skew). As also shown in Table 1, both measures' range of agreement scores is wide. Within-group agreement regarding task significance ranges between 0.18 and 0.58, while within-group agreement regarding group hostility varies from 0.09 to as high as 0.56. At this point, and given the wide quasi-confidence intervals for each of the measured constructs, the obvious question may be: "So how do I interpret these estimates?" We view this as a very important question, and one that we thoroughly address in the Discussion. Briefly, however, r_{WG} (for single item measures) and $r_{WG(J)}$ (for multiple-item measures) are only two of many complementary indices that need to be estimated and judgments should be based on the overall magnitude and pattern of results.

7.4. Summary

On the basis of this empirical example, we hope to have shed light on the use (and potential for misuse) and application of interrater agreement indices; specifically, the most popular estimates of r_{WG} and $r_{WG(J)}$ (James et al., 1984). Although our results suggested that clustering effects were present in all three variables, only leadership climate exceeded the 0.70 threshold that is frequently (and inappropriately) invoked (Lance et al., 2006). Nevertheless, this latter finding should also be viewed with caution, since the $r_{WG(J)}$ estimates were derived using the rectangular (uniform) null distribution. In sum, when attempting to decide if data should be aggregated, we urge leadership researchers to carefully consider the potential consequences of their data aggregation decisions.

8. Discussion

Despite advances in multilevel leadership theory, research, and methodologies, an integrative framework has yet to be offered for conducting aggregation analyses and reporting such research. Consequently, there is considerable variation in researchers' justification of data aggregation. To illustrate, consider the following statements extracted from articles in high-quality journals. Note that these statements are common and we see no need to single out particular authors; however, citations are available on request:

> "The r_{WG} mean value for the leadership scores for SSN level was 0.82 (ICC1 = .20, ICC2 = 0.57) and mean r_{WG} for leadership scores for NOs was 0.83 (ICC1 = .24, ICC2 = 0.76). Although no absolute standard value for aggregation based on r_{WG} and ICC have been established, an r_{WG} equal to or greater than 0.70 and ICC(1) values exceeding 0.05 (Bliese, 2000) is considered sufficient to warrant aggregation. Based on the results, we concluded that it was statistically appropriate to assess transformational leadership as a group-level variable." (2004, *Journal of Organizational Behavior*).

> "With respect to aggregation, there was evidence that transformational leadership varied significantly across stores, $F(67, 379) = 1.59$, $p < .01$. Intraclass correlation ICC(1), ICC(2), and median $r_{WG(J)}$ values were .08, .37, and .95. Transactional leadership also varied significantly across stores, $F(67, 379) = 1.73$, $p < .01$. ICC(1), ICC(2), and median $r_{WG(J)}$ values were .10, .42, and .80. Passive leadership significantly varied across groups as well, $F(67, 380) = 1.78$, $p < .01$. ICC(1), ICC(2), and median $r_{WG(J)}$ values were .10, .43, and .85 … In light of all the evidence regarding the ANOVA, ICC(1), ICC(2), and $r_{WG(J)}$, we proceeded to create aggregate measures of transformational, transactional, and passive leadership." (2005, *Journal of Applied Psychology*).

Table 1
Within-group agreement statistics.

	$r_{WG(J).uniform}$	$r_{WG(J).measure\text{-}specific}$		
Measure	Mean	Shape	σ_E^2	Mean
Leadership climate	.87	Slight skew	1.34	.41
Leadership climate	–	Normal	1.04	.10
Task significance	.58	Slight skew	1.34	.18
Group hostility	.56	Moderate skew	0.90	.09

Notes. $r_{WG(J)}$ is reported because multi-item measures were used. Shape = the shape of an alternative null distribution; σ_E^2 = variance of an alternative null distribution. Variance estimations for measure-specific null distributions (i.e., slight skew, normal, and moderate skew) were taken from LeBreton and Senter (2008, p. 832).

In contrast, a third example successfully argued that:

> "We calculated the intraclass correlations (ICCs; Bliese, 2000) and the within-group agreement (r_{WG}; James et al., 1984). The average r_{WG} was .64, ranging from .45 to .71, whereas the ICCs were as follows: ICC(1) was .10 and ICC(2) was .60. The group effect (i.e., the F value for the ANOVA) was significant at $p = .05$. Although these statistics suggest some group-level effects, we decided to treat transformational leadership at the individual follower level. Our decision was based in part on the r_{WG} value falling below the traditional cutoff recommended for forming groups of .70, the ICC(1) value being relatively low, as well as based on the individual level of analysis used for our intervening and performance outcomes." (2008, *Personnel Psychology*).

We find it interesting that all three examples focused on transformational leadership and even used the same 20-item measure, and yet they differed in how key data aggregation analyses were reported. Furthermore, the third example ignores the group-level effects due to a mean r_{WG} value below the "traditional cutoff" (an invalid assumption; Lance et al., 2006), and despite a statistically significant ANOVA F-statistic and ICC(1) and ICC(2) values that were in line with the others' research. This latter decision to overlook group-level effects is potentially problematic, given that researchers may draw erroneous conclusions (due to biased standard errors) when unit membership is a known source of variance but is excluded from statistical analyses (Bliese & Hanges, 2004). Hence, we maintain that such decision inconsistencies can have important theoretical and practical consequences; most notably, confused readers as well as the potential for different conclusions based upon empirical results.

In an effort to provide consistency in the multilevel leadership arena, we have developed a brief list of critical decisions that should be considered when designing a multilevel study, aggregating lower level data to a higher level of analysis, and reporting on its findings. Our goal in offering this list is to encourage researchers to consider the potential implications of their data aggregation decisions when conducting leadership research. What follows may also be of use to journal referees commissioned to review submitted manuscripts on aggregate leadership phenomena.

9. Step 1: select theoretically defensible null distributions

To calculate r_{WG} and $r_{WG(J)}$, the expected variance that assumes no agreement among raters must be computed; this estimate is based on a null distribution reflecting a total lack of agreement. Although this may seem like a simple task, methodologists have noted that "choosing the null distribution is the single greatest factor complicating the use of r_{WG}-based indices" (LeBreton & Senter, 2008, p. 829). Although the vast majority of researchers have invoked the rectangular (uniform) null distribution, there are a number of alternative null distributions available to researchers. Further, James et al. (1984) and, more recently, LeBreton and Senter (2008), have explicitly recommended that researchers use a small but inclusive set of null distributions when computing r_{WG}-based indices. When deciding on which null distributions are the most appropriate, we encourage researchers to consider multiple sources of information. Essentially, these efforts should produce the "identification of several possible nulls, perhaps often, the uniform distribution" (James et al., 1984, p. 94). Then, based on the evidence obtained, the informed researcher can compute a range of r_{WG}-based estimates, thereby increasing the likelihood of the true estimate falling within this range of scores.

9.1. A rectangular null distribution

Unless there is a priori knowledge of response bias, the rectangular (uniform) null distribution may be "[t]he most natural candidate to represent nonagreement," because it assumes that all answering options have the same probability of being selected by the rater (Cohen et al., 2009, p. 149). It is well-known that the rectangular distribution produces inflated values of r_{WG} and $r_{WG(J)}$ (as it yields large error variance estimates); therefore, this estimate should be considered an upper-bound of within-group agreement.

9.2. Identify alternative null distributions

We recommend that researchers do their homework, in the sense that they should gather as much information as possible on the possibility of theoretically defensible, alternative forms of the null distribution. In many instances, one's theoretical rationale for a proposed alternative null can be supplemented with empirical evidence gathered from prior studies employing the same measure(s). Using past empirical research to identify a more realistic null distribution is in line with James et al.'s (1984) initial recommendations. Researchers can use the observed distributions (i.e., the variance) from published research (using the same measure) and/or additional data (but not the focal data) to compute a small but inclusive set of alternative null distributions to be used in the analyses (see, e.g., Kozlowski & Hults, 1987). By incorporating one (or more) alternative null distributions when estimating within-group agreement, researchers create a lower bound estimate for r_{WG} and $r_{WG(J)}$.

For example, assume researchers are interested in aggregating subordinates' ratings of managers' abusive supervisory behavior by means of the measure developed by Tepper (2000). According to Tepper, abusive supervision refers to "subordinates' perceptions of the extent to which supervisors engage in the sustained display of hostile verbal and nonverbal behaviors, excluding physical contact" (p. 178). Thus, abusive supervision researchers have reason to anticipate a moderate to large triangular distribution (i.e., a quasi-normal distribution), wherein a higher proportion of subordinates use the middle (e.g., 3 = neutral) response option to evade the question set. Alternatively, these researchers, having done their due diligence, might anticipate a moderate to

large positive leniency response bias, as prior studies have acknowledged that abusive supervision is a low base rate phenomenon, in which case a moderate to large skew might be more appropriate. In the present scenario, each possibility appears equally plausible and, thus, they may wish to compute r_{WG}-based estimates using the rectangular distribution, as well as both alternative null distributions.

9.3. Caveat to Step 1

The selection of the appropriate set of null distributions to estimate a range of r_{WG} or $r_{WG(J)}$ values should be based on theory (see, e.g., Lebreton et al., 2003). If theory is seemingly unavailable, researchers *should never* employ the observed distribution to propose a hypothesized null distribution. Given that true scores are often confounded with systematic rating biases, a particular observed distribution may reflect true scores or it may be (partially or fully) attributed to other factors (e.g., response bias). As James et al. (1984) noted, this possibility "*underscores the need to obtain evidence other than the observed distribution to propose nulls. This other evidence consists of the aforementioned use of knowledge from prior research*" (italics added, p. 94).

10. Step 2: calculate interrater agreement

We recommend using two (the minimum) or three null distributions when computing r_{WG} and $r_{WG(J)}$ estimates. As previously discussed, researchers may wish to use the rectangular (uniform) distribution to obtain an upper bound estimate, and a more "realistic" measure-specific null distribution to compute a lower bound estimate. On the basis of this strategy, researchers can calculate a range of r_{WG}-based estimates (i.e., quasi-confidence intervals) within which the true estimate is more likely to fall (James et al., 1984).

10.1. Interpreting interrater agreement

Traditionally, 0.70 has been used as a cut-point for establishing high versus low interrater agreement. Lance et al. (2006) have noted that the 0.70 cut-point is a frequently and inappropriately applied heuristic, and that, interestingly, this threshold is often attributed to James et al. (1984). In tracing this widely used heuristic to its (alleged) original source, Lance et al. found that what "James et al. (1984) actually said regarding the .70 cutoff criteria for r_{WG} was … nothing" (2006, p. 207). Nevertheless, despite doubts raised regarding the usefulness of the absolute 0.70 standard, we note that leadership researchers continue to rely on it. For example, in a 2008 study published in the *Journal of Applied Psychology*, it was reported that (citation available upon request):

> "The r_{WG} for transformational leadership was .96, r_{WG} for support for innovation was .95, and r_{WG} for climate for excellence was .89. All the r_{WG} values were above the critical cutoff value of .70 (James et al., 1984) and thus suggested that it was appropriate to aggregate individual responses to the team level. However, as the r_{WG} has been criticized for using a uniform distribution, we also calculated the a_{WG} index."

The above illustration is also a reminder that researchers (and not the r_{WG} agreement statistic per se) choose to utilize a rectangular (uniform) null distribution when computing r_{WG}-based indices. Thus, it would seem that as a field, we need to be more aware of all the issues pertaining to r_{WG}-based indices and how best to interpret them.

Furthermore, contemporary thinking is that the 0.70 threshold "artificially dichotomizes agreement in a manner that is inconsistent with James et al.'s (1984) original intention, and it may not be useful for justifying aggregation" (LeBreton & Senter, 2008, p. 835). That is, rather than drawing an arbitrary "line in the sand," a researcher should consider interrater agreement in terms of: "lack of agreement" = .00 to .30; "weak agreement" = .31 to .50; "moderate agreement" = .51 to .70; "strong agreement" = .71 to .90, and; "very strong agreement" = .91 to 1.00 (LeBreton & Senter, 2008; see also Brown & Hauenstein, 2005). Cut-points used to justify the aggregation of consensus composition models can still be established using this more-inclusive set of standards; however, we suggest that a lower-bound cut-point for denoting high versus low agreement should correspond to the theoretical expectations and previously published evidence (when available) and, not, an absolute standard (> 0.70). Moreover, a research study's purpose and ever present practical considerations would also require consideration when values used to justify aggregation are determined.

10.2. The statistical significance of r_{WG}-based indices

Multilevel leadership researchers may also be interested in addressing the question: "Do the interrater agreement values in my study sufficiently differ from chance agreement?" There are statistical significance tests for evaluating r_{WG} and $r_{WG(J)}$ values against a null hypothesis. Although these tests cannot resolve the issue of how large an estimate must be to justify aggregation, we find them helpful. These significance tests allow researchers to address a necessary precondition associated with consensus composition models; that is, rejection of the null hypotheses of no agreement. As Cohen et al. (2009) observed, such tests "do indicate that some agreement exists, regardless of its magnitude" (p. 151). And yet, like any significance test, statistical power needs to be considered. In other words, one can observe a high r_{WG} value (e.g., 0.75) in a small group size that is not statistically significant, and a relatively moderate value (e.g., 0.50) in a large group size that is statistically significant. Researchers interested in learning more about the statistical significance of r_{WG} and $r_{WG(J)}$ are referred to Cohen et al. (2009) and Dunlap et al. (2003).

Recently, Pasisz and Hurtz (2009) introduced a novel approach that tests for differences between two or more groups' within-group agreement. They provide a test that directly compares the difference between two or more r_{WG} or $r_{WG(J)}$ estimates. This procedure is an important step forward for leadership researchers regularly computing r_{WG} or $r_{WG(J)}$ estimates, in that it allows them to contrast r_{WG}-based estimates from a current dataset with r_{WG}-based estimates from published research. Pasisz and Hurtz's (2009) approach is also capable of directly comparing r_{WG} or $r_{WG(J)}$ values derived from a single sample. If certain large groups are found to yield low agreement, researchers could conduct exploratory post hoc searches for potential moderators (e.g., tenure of the work unit leader). Although additional research is needed to explore the power and robustness of these tests in real-world contexts (e.g., varying group sizes, total number of group sampled), we anticipate Pasisz and Hurtz's approach will be advantageous when exploring interrater agreement within groups.

10.3. Options when some (but not all) groups have high interrater agreement

It is very likely that any given data set will have groups with high agreement and low agreement. In such situations, one option is to eliminate groups with low levels of agreement prior to analyzing the data (i.e., hypothesis testing). We advise researchers against this first alternative because losing valuable data points is never ideal; statistical power will be reduced if groups with low r_{WG} or $r_{WG(J)}$ values are excluded. A second and more viable option is to conduct a series of sensitivity analyses, wherein researchers analyze the data with and without the identified low agreement groups. If the sensitivity analyses yield a similar pattern of results, researchers can be more confident that the mixing of high and low agreement groups was not a serious enough problem to disparage hypothesis testing. If the results are inconsistent, researchers also have options. For low agreement groups, it may, for example, be possible to move from the focal unit of analysis to a subgroup level. Researchers interested in learning more about these topics are referred to Pasisz and Hurtz (2009) and Gooty and Yammarino (2011).

10.4. Caveat to Step 2

There is a disincentive for researchers to compute r_{WG} and $r_{WG(J)}$ estimates using any alternative null distribution — the interrater agreement values will, in most instances, be lower in magnitude than the r_{WG}-based estimates computed using a rectangular (uniform) null distribution. LeBreton and Senter (2008) may state it best: "we acknowledge this disincentive but challenge researchers to use sound professional judgment when choosing the null distributions to estimate r_{WG} and challenge reviewers to hold authors accountable for the decisions they make involving null distributions" (p. 836). We could not agree more.

10.5. When are r_{WG}-based indices (less) useful?

In adopting either a *direct consensus* or *referent-shift consensus* composition model, it is assumed that members of each group share a common perception with regard to a target construct and that any observed variability in members' ratings is noise or measurement error. Accordingly, high interrater agreement is a necessary prerequisite for the aggregation of lower level data to a higher level because agreement indicates consensus. For this purpose, r_{WG} and $r_{WG(J)}$ indices are indeed useful.

Whereas consensus composition models have been the dominant theoretical basis for explaining the emergence of aggregate leadership phenomena, researchers (e.g., Cole, Bedeian, & Bruch, 2011; Hooper & Martin, 2008) have also applied an alternative approach termed a *dispersion composition* model. In contrast to consensus composition models, a *dispersion composition* model conceptualizes within-group variance as a focal construct of theoretical importance, rather than as a statistical prerequisite for aggregation (Chan, 1998). In these instances, within-group agreement (e.g., r_{WG}-based) cut-points are less important for justifying aggregation (Cole et al., in press). In fact, r_{WG} and $r_{WG(J)}$ values are used to index dispersion variables (Roberson, Sturman, & Simons, 2007) and, thus, provide redundant information.

In addition, r_{WG} and $r_{WG(J)}$ indices evaluate within-group rater agreement and do not consider between-group variability (see Eq. 1). Nevertheless, it is widely acknowledged that between-group variability has important implications for multilevel leadership studies. For example, it is possible for members of each group to agree, but for groups to exhibit little or no between-group variance. In this scenario, and despite high interrater agreement, the aggregated variable will be of little predictive value, because a lack of reliable mean differences substantively reduces the aggregate variable's explanatory power (Biemann & Heidemeier, 2010). Thus, in the absence of between-group variability, the aggregate or group-level construct's validity may be questioned (Chan, 1998).

A reason for the lack of between-group variance in a data set may lie in researchers' sampling strategy, which can greatly influence the variability both within and between groups. Assume, for example, that a researcher collected ratings of supervisors' transformational leadership behavior from 500 followers nested in 100 work teams from one organization in a single industry. In a follow-up study, the researcher collected data on the same measure from the same number of followers and teams, but the teams were nested in 20 organizations in four industries. In the first study, the researcher may not find large between-group differences, because the teams belong to the same organization. In contrast, between-group (i.e., team-level) variability is likely to be larger in the second study than in the first, because teams in the second study were drawn from a more heterogeneous population (see, e.g., Chen, Mathieu, & Bliese, 2004; George, 1990). Thus, in addition to a r_{WG}-based estimate that evaluates rater consensus within each group or unit, leadership researchers should also consider omnibus estimates that apply across groups (which we describe next), thereby ensuring that the aggregated variable varies both within and between units of analysis.

11. Step 3: calculate intraclass correlation coefficients (ICCs)

At this point, mounting concerns regarding r_{WG} and $r_{WG(J)}$ suggest that they should not be used as the sole index to justify aggregating lower level data (based on consensus composition models) to a higher level of analysis. Thus, what can multilevel leadership researchers do to test the appropriateness of their conceptualizations of lower level variables as higher level aggregates? Along with LeBreton et al. (2003), we suggest that researchers should examine both the interrater agreement (i.e., r_{WG}-based indices) and interrater reliability (i.e., ICCs) statistics to provide a form of "psychometric checks and balances" concerning interrater similarity (p. 121).

Interrater agreement (IRA) emphasizes the interchangeability between judges' ratings, whereas interrater reliability (IRR) emphasizes the relative consistency in multiple judges' ratings of multiple targets (Kozlowski & Hattrup, 1992; Tinsley & Weiss, 1975). Although both concepts address the similarity of judges' ratings, they differ in how they go about determining interrater similarity (Lüdtke & Robitzsch, 2009). Thus, it is not unreasonable to invoke both IRA and IRR indices when attempting to justify aggregation. The most common IRR indices are intraclass correlation coefficients — particularly ICC(1) and ICC(2) (Bartko, 1976; Shrout & Fleiss, 1979). Both forms of ICCs can be calculated from a one-way random-effects ANOVA, where the variable of interest (e.g., followers' ratings of transformational leadership) is the dependent variable and group membership is the independent variable. In doing so, a researcher uses ICCs to ensure that there is sufficient variance within and between units of analysis. Alternatively, the notion that variability among group members' judgments may provide meaningful information is consistent with the logic underlying within and between analysis (WABA; Dansereau & Yammarino, 2000). Readers interested in the WABA procedure should consult Castro (2002), Dansereau and Yammarino (2000; 2006), and Dansereau, Cho, and Yammarino (2006).

ICC(1) demonstrates the amount of variance in a variable that is attributable to group membership and is calculated as the ratio of between-group mean square (MSB) variance to total variance (sum of MSB and within-group mean square [MSW] variance). For example, an ICC(1) = MSB/(MSB + MSW) = .06 suggests that group membership explains six percent of the variance in individual group-members' ratings. Consequently, ICC(1) is typically considered an estimate of effect size (see Bliese, 2000). If the ICC(1) is statistically different from zero, there is evidence to justify making the group the focal unit of analysis (Chen et al., 2004). ICC(2) assesses the reliability of the group-level means, indicating how reliably the aggregate mean rating (across group members) distinguishes between groups (Bliese, 2000). Bliese (1998, 2000) has suggested ICC(2) provides evidence of emergent properties and is calculated using MSB − MSW/MSB. Given that ICC(1) and ICC(2) are based on variance partitioning, the underlying assumptions of ANOVA must be met when calculating either index. These assumptions include approximate equal-interval measurement, normally distributed group scores, independent between-group observations, and homogeneity of variances within groups (for details, see Dansereau & Yammarino, 2006; McGraw & Wong, 1996).

11.1. Interpreting interrater reliability

Mirroring r_{WG} and $r_{WG(J)}$, there are no definitive rules when attempting to determine the ICC values necessary to justify aggregation. Whereas LeBreton and Senter (2008) have suggested that an ICC(1) = .05 represents a small to medium effect (p. 838), Bliese (1998) has simulated conditions where only 1% of the variance is attributed to group membership (ICC(1) = .01) and, still, strong group-level relationships were detected that were not evident in the lower level data. Echoing this prior work, we believe that researchers should set a priori cut-points for the ICCs they believe are suitable for their specific research question and study context. We suspect that in many instances researchers can use published empirical research to set lower bounds for the ICCs. Researchers interested in learning more about computing and interpreting ICCs in a multilevel context are referred to Bliese (1998, 2000) and LeBreton and Senter (2008).

Although there is nothing that inherently requires leadership researchers to report both IRA and IRR estimates in a single study, ICC(1) and ICC(2) are established complements to r_{WG}-based indices. On this basis, we encourage multilevel leadership researchers to consider reporting all three indices when *direct consensus* or *referent-shift consensus* composition models are used to specify functional relationship between leadership phenomena across different levels of analysis. By computing IRA and IRR statistics and comparing these values to cut-points (a priori) based on previously published research, multilevel leadership researchers can better determine if their data "lack reliability, agreement, neither, or both" (LeBreton & Senter, 2008, p. 840). Consequently, the ultimate goal should be to obtain empirical evidence that allows for well-informed aggregation judgments based on the overall magnitude and pattern of these complementary indices.

11.2. Caveat to Step 3

Using simulation methods, Beal and Dawson (2007) have shown that Likert-type response formats (most often used by leadership research) can adversely influence ICC estimates. Specifically, the use of Likert scales with common response formats (i.e., 5 points) were shown to substantially underestimate ICC(1); further, ICC(2) and group-level correlations were also underestimated, but to a lesser extent. Multilevel leadership researchers may wish to consider Beal and Dawson's findings when deciding whether aggregating lower level data to a higher level is justified. For those planning a future multilevel study, Beal and Dawson found that using response formats with a larger number of options (i.e., 7 points or 9 points) significantly reduces the adverse effects of Likert-type response ramps on ICC estimates.

12. Advice for the reporting of data aggregation

Leadership researchers need to be more specific in their reporting of data aggregation. The following text was extracted from a 2008 study published in *The Leadership Quarterly* (citation available upon request):

> "Since there were multiple raters of CEO leadership as well as the moderating variables, we first tested within-company variance using James et al. (1984) r_{WG} procedure. We found at least 90% of the companies had a r_{WG} value of .7 or higher for all of the scales. Interrater agreement based on intraclass correlations also showed acceptable ranges for all of the measured variables (transformational leadership = .88; empowerment = .78; climate for innovation = .84; centralization = 74; formalization = .85; competition = .74; and uncertainty = .71). Based on these results, we aggregated our data to the company level and conducted all subsequent data analyses at this level."

On reviewing this example, the reader is left to speculate on which agreement index, r_{WG} or $r_{WG(J)}$, was used. On the surface, it appears that the r_{WG} index was applied and, yet, the constructs assessed were based on multi-item measures. This study also neglects to report the type of null distribution used to compute the interrater agreement statistics (e.g., rectangular [uniform] null, slightly skewed, or triangular distribution). The quoted text also appears to confuse interrater agreement (i.e., r_{WG}) with interrater reliability and associated intraclass correlation coefficients (ICCs). Finally, the study reports, for example, that the ICC for transformational leadership was .88. This latter finding cannot be unambiguously interpreted, as the study neglects to report if this ICC value is reflective of the amount of variance attributable to group membership (namely, ICC[1]) or assesses the reliability of the company-level mean (namely, ICC[2]).

Our argument is simply that multilevel leadership researchers should report every decision taken when determining if aggregation is justified, and should describe their reasoning behind each decision so that the end users (e.g., *LQ* readers) can fully evaluate the quality of the aggregated data. To illustrate this point, we conducted a manual search of the James et al. (1984) article in 13 business journals known to publish leadership and group-based research in the period 2003 to 2008.[3] This search resulted in 190 articles. Of these, 169 articles reported a total of 548 r_{WG} or $r_{WG(J)}$ estimates. A majority of these studies (55.1%) reported the overall mean value of the r_{WG}-based estimate; however, the median was reported in 24.8% of the studies and no decision was documented in 20.1% of the cases. Moreover, interrater reliability indices, ICC(1) and ICC(2), were reported in only 53.2% and 35.6% of the studies, respectively. These findings demonstrate that a large number of studies omitted potentially important information concerning data aggregation.

Mirroring the aforementioned pattern of results, leadership researchers have likewise provided inconsistent information regarding data aggregation. For example, from this database of 169 articles (i.e., citing James et al., 1984), we identified studies assessing individual followers' ratings of their managers' transformational leadership behavior, and then proceeded to aggregate these data to a higher level of analysis. We identified a total of 17 empirical articles that assessed transformational leadership and also explicitly cited the work by James et al. (see Table 2). Within this subsample, nine studies reported the overall mean r_{WG}-based value, two reported the median, one reported both the mean and median, and another five studies offered no information as to whether the mean or median was calculated and reported. Further, 14 of the 17 studies ignored previously noted concerns, and computed r_{WG}-based estimates using only the rectangular (uniform) null distribution. The remaining three studies reported $r_{WG(J)}$ values based solely on a slightly skewed null distribution. The ICC(1) index was reported in 11 of the 17 studies and the ICC (2) index in 10 of the 17 instances. Due to the wide variation in reporting on data aggregation results, only three of the 17 studies provided interrater agreement (i.e., $r_{WG(J)}$) and interrater reliability (i.e., ICC[1] and ICC[2]) estimates that were comparable across studies. This may have occurred because it was believed that it does not matter which aggregation statistics are reported, so there was no reason to report all of this information; or the whole issue of interrater agreement and reliability may merely have been overlooked. In adopting a *direct consensus* or *referent-shift consensus* model, we take the position that it is important for the reader to be able to evaluate one's data aggregation decisions, and this requires researchers to report both interrater agreement and reliability indices.

In an attempt to encourage future leadership research to take up this challenge, we have created an example table (see Table 3) that others should feel free to use and modify. As depicted, the data aggregation table might provide relevant information regarding the r_{WG}-based estimates based on a rectangular (uniform) distribution and one or two alternative, measure-specific null distributions; that is, a small but inclusive set (see James et al., 1984). The selection of an appropriate alternative null was detailed earlier, with the choice reported in the table as the "shape" of the null distribution. We likewise recommend that the variance of the alternative (measure-specific) null distribution (σ^2_{EU} in Table 3) be provided. For example, in the handful of studies that did apply an alternative null, researchers most often described the form of the null distribution (e.g., "slightly skewed" or "moderately skewed") used. Unfortunately, it is impossible to compare or replicate these findings, because there are many variance distributions that match a "slightly" or "moderately" skewed description. Only the variance of the null distribution is needed to calculate r_{WG}-based estimates. Thus, reporting the variances associated with alternative null distributions (σ^2_{EU} in Table 3) will not only make aggregation results more transparent, but also more comparable across empirical studies. Further, as also shown in Table 3, we suggest researchers report standard deviations associated with the r_{WG}-based estimates.

[3] The journals included: *Academy of Management Journal*, *Administrative Science Quarterly*, *Journal of Applied Psychology*, *Journal of International Business Studies*, *Journal of Management*, *Journal of Occupational and Organizational Psychology*, *Journal of Organizational Behavior*, *The Leadership Quarterly*, *Management Science*, *Organization Science*, *Organizational Behavior and Human Decision Processes*, *Personnel Psychology*, and *Small Group Research*.

Table 2
Empirical studies that aggregated followers' ratings of transformational leadership to a higher level (2003–2008).

Article	Decision rules for data aggregation (paraphrased)	r_{WG}			ICC (1)	ICC (2)	Aggregated the data?
		r_{WG} reported?	Null distribution	Reported aggregate			
Bass et al. (2003)	Between 70% and 80% of the r_{wg} values for all survey scales fell above the .70 cutoff suggested by James et al. for aggregating ratings from an individual to a group level of analysis.	✓	Uniform	Mean	?	?	Yes
Bono and Judge (2003)	Aggregating leadership reports across followers was deemed justifiable in these data by a significant ICC(1) value and an ICC(2) value. An average r_{WG} across groups (assuming a slight negative skew in the data) further supports aggregation.	✓	Slight skew	?	✓	✓	Yes
Avolio, Zhu, Koh, and Bhatia (2004)	Although no absolute standard value for aggregation based on r_{WG} and ICC have been established, an r_{WG} equal to or greater than 0.70 and ICC1 values exceeding 0.05 (Bliese, 2000) are considered sufficient to warrant aggregation.	r^*_{WG}	Uniform	Mean	✓	✓	Yes
Berson and Avolio (2004)	The r_{WG} values for TFL were above James et al.'s (1984) recommendation, ranging from .71 to .89 for all departments, with an average of .84.	✓	Uniform	Mean	?	?	Yes
Bommer, Rubin, and Baldwin (2004)	To check whether the raters were "'seeing the same thing,"' we calculated a measure of rater agreement (i.e., r_{WG}). James et al. (1984) assert that the r_{WG} statistic provides a valid estimate of actual within-group agreement, and that an r_{WG} equal to or greater than .70 demonstrates acceptable levels of agreement and suggests that aggregation to the group level is valid.	✓	Uniform	Mean	?	?	Yes
Hofmann and Jones (2005)	The analysis of variance (ANOVA), ICC(1), and $r_{WG(J)}$ values is in keeping with past research involving aggregation … In light of all the evidence regarding the ANOVA, ICC(1), ICC (2), and $r_{WG(J)}$, we proceeded to aggregate …	✓	Uniform	Median	✓	✓	Yes
Liao and Rupp (2005)	Following James et al. (1984) and Kozlowski and Hults (1987), we assessed interrater agreement by computing James et al.'s $r_{WG(J)}$, which adjusted for a slight negative skew in the expected variance. We then conducted one-way analyses of variance and found significant between-groups variance for all variables. We further obtained the intraclass correlation (ICC1) and reliability of group mean (ICC2) values. These values are comparable to the median ICC values of aggregated constructs reported in prior studies of TFL.	✓	Slight skew	Mean	✓	✓	Yes
Rubin et al. (2005)	Following George's (1990) suggestion and previous TFL research, we computed a measure of within-group agreement (r_{WG}). … less than 1% of the r_{WG} calculations fell below the acceptable .70 cutoff.	✓	Uniform	Mean	?	?	Yes
Bono et al. (2007)	This procedure [data aggregation] was consistent with past research and was deemed justifiable in these data under James et al.'s (1993) recommendations for data assumed to have a slight negative skew. Although no absolute standard for aggregation based on the ICC(1) or r_{WG} has been established, Bliese (2000) reported values in organizational research from .05 to .20 and rare instances of values exceeding .30. Typically, r_{WG} values greater than .70 are used to justify aggregation.	✓	Slight skew	Mean	✓	✓	Yes
Rowold and Heinitz (2007)	Following the recommendations made by McGraw and Wong (1996), interrater agreement (ICC1 and ICC2) and within-group agreement indices (r_{WG}) were calculated. Table 5 indicates that the raters highly agreed on the three leadership scales.	✓	Uniform	Mean	✓	✓	Yes
Shin, Morgeson, and Campion (2007)	In sum, these results met or exceeded the levels of reliability and agreement found in previous research dealing with aggregation issues (e.g., Campion et al., 1993). Thus, aggregating the responses to the team level was appropriate.	✓	Uniform	Mean and median	✓	✓	Yes
Eisenbeiss, van Knippenberg, and Boerner (2008)	The r_{WG} values were above the critical cutoff value of .70 (James et al., 1984) and thus suggested that it was appropriate to aggregate individual responses to the team level.	✓	Uniform	?	?	?	Yes

(continued on next page)

Table 2 (*continued*)

Article	Decision rules for data aggregation (paraphrased)	r_{WG}			ICC (1)	ICC (2)	Aggregated the data?
		r_{WG} reported?	Null distribution	Reported aggregate			
Herold, Fedor, Caldwell, and Liu (2008)	We computed r_{WG} values to examine agreement among group members. Next, we computed intraclass correlation coefficients (ICC1) to examine within and between-groups variance in leader assessments (Bliese, 2000). The median r_{WG} values … indicate strong agreement about each leader's transformational leadership behaviors. ICC1s were high (.10; Bliese, 2000), indicating significant between-groups variance …	✓	Uniform	Median	✓	?	Yes
Jung, Wu, and Chow (2008)	We found at least 90% of the companies had an r_{WG} value of .7 or higher for all of the scales.	✓	Uniform	?	?	?	Yes
Kearney (2008)	The mean inter-rater agreement value (r_{WG}) as well as the intra-class correlation coefficients confirmed that this was the case and that averaging responses to the team level was justified (Bliese, 2000).	✓	Uniform	Mean	✓	✓	Yes
Luria (2008)	Glick (1985) argued that aggregation of individual responses requires an above-threshold consensus ($r_{WG} = 0.70$, assuming uniform null distribution).	?	?	?	✓	✓	Yes
Walumbwa, Avolio, and Zhu (2008)	We calculated the intraclass correlations (ICCs; Bliese, 2000) and the within-group agreement (r_{WG}). The average r_{WG} was .64, ranging from .45 to .71, whereas the ICCs were as follows: ICC(1) was .10 and ICC(2) was .60. The group effect (i.e., the F value for the ANOVA) was significant at $p = .05$. Although these statistics suggest some group-level effects, we decided to treat TFL at the individual follower level. Our decision was based in part on the r_{wg} value falling below the traditional cutoff recommended for forming groups of .70, the ICC (1) value being relatively low, as well as based on the individual level of analysis used for our intervening and performance outcomes (Rousseau, 1985).	✓	Uniform	?	✓	✓	No

Finally, as likewise shown in Table 3, researchers should report *F* ratios, ICC(1), and ICC(2) values. The *F* ratio is the result of an ANOVA-based significance test of between-group differences and indicates the statistical significance of group membership. For additional ideas on how to summarize and report data aggregation statistics, we refer the reader to Rowold and Heinitz (2007) and Walker et al. (2008). To ease calculations, we have created an Excel-based statistical tool that computes interrater agreement statistics and all complementary indices shown in Table 3. It can be downloaded from Michael S. Cole's website (www.sbuweb.tcu.edu/mcole).

13. Conclusion

To advance contemporary leadership literature, multilevel frameworks are increasingly used to test theory and establish empirical findings. Nevertheless, the soundness of such work depends on how researchers tackle critical aggregation questions. We observe, as have others (Castro, 2002; Cohen et al., 2001; Lance et al., 2006; LeBreton & Senter, 2008), that r_{WG}

Table 3
A template to report aggregation results for consensus composition models[a].

Measure	$r_{WG(J).uniform}$		$r_{WG(J).measure\text{-}specific}$						
	Mean	SD	Shape	σ_E^2	Mean	SD	F ratio	ICC(1)	ICC(2)
Empowering leadership climate (5)[b]	.91	0.33	Slight skew	1.34	.60	0.31	3.09**	.06	.50
Team thriving (7)[c]	.75	0.30	Triangular	2.10	.31	0.19	4.75**	.10	.75
Team commitment (7)	.87	0.35	Moderate skew	2.14	.09	0.30	5.15**	.09	.65
Team satisfaction (5)	.86	0.21	Moderate skew	0.90	.12	0.25	3.18**	.07	.52
Team satisfaction (5)	–	–	Heavy skew	0.44	.05	0.17	–	–	–

Notes. SD = standard deviation of $r_{WG(J)}$ values; shape = the shape of an alternative null distribution; σ_E^2 = variance of an alternative null distribution. Variance estimations for measure-specific null distributions (i.e., slight skew, normal, and moderate. skew) were taken from LeBreton and Senter (2008, p. 832).

[a] Measures and estimates were invented for illustration purposes only.

[b] Denoting a 5-point scale.

[c] Denoting a 7-point scale.

** $p < .01$.

and $r_{WG(J)}$ values (and other variance-based indices of interrater agreement) can be misleading in that variables with no clustering effect can easily exceed the common cut-point of 0.70, and variables with a significant clustering effect can have values far below 0.70. Hence, adequate attention must be given to the various decisions faced when aggregating followers' ratings of their managers' leadership behavior and be cognizant of those decision's consequences. Further, when assessing within-group agreement, it is important to assume that more than one response distribution exists. Given the availability of expected variance estimates for calculating a small set of null distributions, we see no logical reason for the multilevel leadership literature to continue relying on the rectangular (uniform) null distribution. Along similar lines, LeBreton and Senter (2008) have called for a "moratorium on the unconditional (i.e., unjustified) use of any null distribution, especially the uniform null distribution" (p. 830).

In sum, researchers confronted with the question of whether it is justified to aggregate lower level leadership ratings to a higher level should not solely rely on r_{WG} and r_{WG}-like indices. A more useful approach – in our opinion – is to (a) estimate r_{WG}-based indices using a small but inclusive set of null distributions, (b) test the statistical significance of clustering effects, and (c) compare the ICCs obtained in the primary study to ICCs reported in comparable studies. We hope that the proposed guidelines will not only help, but encourage future leadership researchers to give adequate thought to their decisions on how best to tackle the complex issues involving aggregate leadership phenomena.

References

Avolio, B. J., Zhu, W. C., Koh, W., & Bhatia, P. (2004). Transformational leadership and organizational commitment: Mediating role of psychological empowerment and moderating role of structural distance. *Journal of Organizational Behavior*, *25*, 951–968.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*, 762–765.

Bass, B. M., Avolio, B. J., Jung, D. I., & Berson, Y. (2003). Predicting unit performance by assessing transformational and transactional leadership. *The Journal of Applied Psychology*, *88*, 207–218.

Beal, D. J., & Dawson, J. F. (2007). On the use of Likert-type scales in multilevel data: Influence on aggregate variables. *Organizational Research Methods*, *10*, 657–672.

Berson, Y., & Avolio, B. J. (2004). Transformational leadership and the dissemination of organizational goals: A case study of a telecommunication firm. *The Leadership Quarterly*, *15*, 625–646.

Biemann, T., & Heidemeier, H. (2010). On the usefulness of the ICC(1) and r_{WG} index to justify aggregation decisions. *Best paper proceedings of the Academy of Management Annual Meeting* August 6–10, Montreal, Canada.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, *1*, 355–373.

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

Bliese, P. D., & Halverson, R. R. (1996). Individual and nomothetic models of job stress: An examination of work hours, cohesion, and well-being. *Journal of Applied Social Psychology*, *26*, 1171–1189.

Bliese, P. D., Halverson, R. R., & Schriesheim, C. A. (2002). Benchmarking multilevel methods in leadership: The articles, the model, and the data set. *The Leadership Quarterly*, *13*, 3–14.

Bliese, P. D., & Hanges, P. J. (2004). Being both to liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods*, *5*, 362–387.

Bommer, W. H., Rubin, R. S., & Baldwin, T. T. (2004). Setting the stage for effective leadership: Antecedents of transformational leadership behavior. *The Leadership Quarterly*, *15*, 195–210.

Bono, J. E., Foldes, H. J., Vinson, G., & Muros, J. P. (2007). Workplace emotions: The role of supervision and leadership. *The Journal of Applied Psychology*, *92*, 1357–1367.

Bono, J. E., & Judge, T. A. (2003). Self-concordance at work: Toward understanding the motivational effects of transformational leaders. *Academy of Management Journal*, *46*, 554–571.

Brown, R. D., & Hauenstein, N. (2005). Interrater agreement reconsidered: An alternative to the r_{WG} indices. *Organizational Research Methods*, *8*, 165–184.

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions: A comparison of intraclass correlation coefficients, $r_{WG(J)}$, hierarchical linear modeling, within- and between-analysis, and random group resampling. *The Leadership Quarterly*, *13*, 69–93.

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

Charnes, J. M., & Schriesheim, C. A. (1995). Estimation of quantiles for the sampling distribution of the r_{WG} within group agreement index. *Educational and Psychological Measurement*, *55*, 435–437.

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multilevel construct validation. *Research in Multi-Level Issues: The many Faces of Multi-Level Issues*, *3*, 273–303.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods*, *6*, 297–310.

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.

Cole, M. S., & Bedeian, A. G. (2007). Leadership consensus as a cross-level contextual moderator of the emotional exhaustion-work commitment relationship. *The Leadership Quarterly*, *18*, 447–462.

Cole, M. S., Bedeian, A. G., & Bruch, H. (2011). Linking leader behavior and leadership consensus to team performance: Integrating direct consensus and dispersion models of group composition. *The Leadership Quarterly*, *22*, 383–398.

Dansereau, F., Cho, J., & Yammarino, F. J. (2006). Avoiding the "fallacy of the wrong level": A within and between analysis (WABA) approach. *Group and Organization Management*, *31*, 536–577.

Dansereau, F., & Yammarino, F. J. (2000). Within and between analysis: The varient paradigm as an underlying approach to theory building and testing. In K. J. Klein, & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 425–466). San Francisco: Jossey-Bass.

Dansereau, F., & Yammarino, F. J. (2006). Is more discussion about levels of analysis really necessary? When is such discussion sufficient? *The Leadership Quarterly*, *17*, 537–552.

Dickson, M. W., Resick, C. J., & Hanges, P. J. (2006). Systematic variation in organizationally-shared cognitive prototypes of effective leadership based on organizational form. *The Leadership Quarterly*, *17*, 487–505.

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

Eisenbeiss, S. A., van Knippenberg, D., & Boerner, S. (2008). Transformational leadership and team innovation: Integrating team climate principles. *The Journal of Applied Psychology*, *93*, 1438–1446.

George, J. M. (1990). Personality, affect, and behavior in groups. *The Journal of Applied Psychology*, *75*, 107–116.

Gooty, J., & Yammarino, F. J. (2011). Dyads in organizational research: Conceptual issues and multilevel analyses. *Organizational Research Methods*, *14*(3), 456–483.
Herold, D. M., Fedor, D. B., Caldwell, S., & Liu, Y. (2008). The effects of transformational and change leadership on employees' commitment to a change: A multilevel study. *The Journal of Applied Psychology*, *93*, 346–357.
Hofmann, D. A., & Jones, L. M. (2005). Leadership, collective personality, and performance. *The Journal of Applied Psychology*, *90*, 509–522.
Hooper, D. T., & Martin, R. (2008). Beyond personal Leader-Member Exchange (LMX) quality: The effects of perceived LMX variability on employee reactions. *The Leadership Quarterly*, *19*, 20–30.
James, D. L., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *The Journal of Applied Psychology*, *69*, 85–98.
James, L. R., Demaree, R. G., & Wolf, G. (1993). r_{WG}: an assessment of within-group interrater agreement. *The Journal of Applied Psychology*, *78*, 306–309.
Jung, D., Wu, A., & Chow, C. W. (2008). Towards understanding the direct and indirect effects of CEOs' transformational leadership on firm innovation. *The Leadership Quarterly*, *19*, 582–594.
Kearney, E. (2008). Age differences between leader and followers as a moderator of the relationship between transformational leadership and team performance. *Journal of Occupational and Organization Psychology*, *81*, 803–811.
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 of the work environment. *The Journal of Applied Psychology*, *86*, 3–16.
Kozlowski, S. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *The Journal of Applied Psychology*, *77*, 161–167.
Kozlowski, S. J., & Hults, B. M. (1987). An exploration of climates for technical updating and performance. *Personnel Psychology*, *40*, 539–563.
Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The source of four commonly reported cutoff criteria. *Organizational Research Methods*, *9*, 202–220.
Lebreton, J. M., Burgess, J., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*, 80–128.
LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{WG}, r^*_{WG}, $r_{WG(J)}$, and $r^*_{WG(J)}$. *Organizational Research Methods*, *8*, 128–138.
LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852.
Liao, H., & Chuang, A. C. (2007). Transforming service employees and climate: A multilevel, multisource examination of transformational leadership in building long-term service relationships. *The Journal of Applied Psychology*, *92*, 1006–1019.
Liao, H., & Rupp, D. E. (2005). The impact of justice climate and justice orientation on work outcomes: A cross-level multifoci framework. *The Journal of Applied Psychology*, *90*, 242–256.
Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.* MA: Addison-Wesley Reading.
Lüdtke, O., & Robitzsch, A. (2009). Assessing within-group agreement: A critical examination of a random-group resampling approach. *Organizational Research Methods*, *12*, 461–487.
Luria, G. (2008). Climate strength—How leaders form consensus. *The Leadership Quarterly*, *19*, 42–53.
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.
Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, *130*, 711–747.
Pasisz, D. J., & Hurtz, G. M. (2009). Testing for between-group differences in within-group interrater agreement. *Organizational Research Methods*, *12*, 590–613.
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*, *9*, 564–588.
Rowold, J., & Heinitz, K. (2007). Transformational and charismatic leadership: Assessing the convergent, divergent and criterion validity of the MLQ and the CKS. *The Leadership Quarterly*, *18*, 121–133.
Rubin, R. S., Munz, D. C., & Bommer, W. H. (2005). Leading from within: The effects of emotion recognition and personality on transformational leadership behavior. *Academy of Management Journal*, *48*, 845–858.
Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *The Journal of Applied Psychology*, *74*, 368–370.
Schriesheim, C. A., Cogliser, C. C., & Neider, L. L. (1995). Is it 'trustworthy'? A multiple-levels-of-analysis reexamination of an Ohio state leadership study, with implications for future research. *The Leadership Quarterly*, *6*, 111–145.
Shamir, B., Zakay, E., Breinin, E., & Popper, M. (1998). Correlates of charismatic leader behavior in military units: Subordinates' attitudes, unit characteristics, and superiors' appraisals of leader performance. *Academy of Management Journal*, *41*, 387–409.
Shin, S. J., Morgeson, F. P., & Campion, M. A. (2007). What you do depends on where you are: Understanding how domestic and expatriate work requirements depend upon the cultural context. *Journal of International Business Studies*, *38*, 64–83.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
Tepper, B. J. (2000). Consequences of abusive supervision. *Academy of Management Journal*, *43*, 178–190.
Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology*, *22*, 358–376.
Walker, A. G., Smither, J. W., & Waldman, D. A. (2008). A longitudinal examination of concomitant changes in team leadership and customer satisfaction. *Personnel Psychology*, *61*, 41–59.
Walumbwa, F. O., Avolio, B. J., & Zhu, W. (2008). How transformational leadership weaves its influence on individual job performance: The role of identification and efficacy beliefs. *Personnel Psychology*, *61*, 793–825.
Yammarino, F. J., & Dansereau, F. (2008). Multi-level nature of and multi-level approaches to leadership. *The Leadership Quarterly*, *19*, 135–141.
Yammarino, F. J., Dionne, S. D., Chun, J. U., & Dansereau, F. (2005). Leadership and levels of analysis: A state-of-the-science review. *The Leadership Quarterly*, *16*, 879–919.