Central Tendency and Matched Difference Approaches for Assessing Interrater Agreement

Michael J. Burke
Tulane University
mburke1@tulane.edu

Ayala Cohen
Technion – Israel Institute of Technology
ieayala@technion.ac.il

Etti Doveh
Technion – Israel Institute of Technology
ierde01@technion.ac.il

Kristin Smith-Crowe
Boston University
kscrowe@bu.edu

Author Note

Michael J. Burke, Management Area, Freeman School of Business, Tulane University; Ayala Cohen and Etti Doveh, Statistics Laboratory, Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel; Kristin Smith-Crowe, Organizational Behavior Department, Questrom School of Business, Boston University.

This research was supported by a grant from the U.S. - Israel Binational Science Foundation. In addition, we would like to thank Patricia Ward, Boston University, for assistance with clerical aspects of our write-up.

Correspondence concerning this article should be addressed to Michael Burke, Management Area, Freeman School of Business, Tulane University, 7 McAlister Pl., New Orleans, LA 70118. Email: mburke1@tulane.edu

Abstract

In Study 1 of this two-part investigation, we present a "central tendency approach" and procedures for assessing overall interrater agreement across multiple groups. We define parameters for mean group agreement and construct bootstrapped confidence intervals around the mean population parameters for r_{WG}, AD, and ICC(1). In Study 2, we extend assessments of overall interrater agreement by developing a "matched difference approach" and procedures for assessing real versus pseudo agreement in a sample of groups. Here, we use random group resampling and the matched difference between assessments of the respective r_{WG}, AD, and ICC(1) values for actual and pseudo groups, with the establishment of bootstrapped confidence intervals around such differences. In both studies, we employ simulated and real data to demonstrate the accuracy and practical utility of the new procedures for assessing agreement with respect to groups. Notably, to generate simulated data for Studies 1 and 2, we developed a new underlying model for multi-level data and procedure for data generation, and we discuss its potential utility for enhancing research in group-level studies. Moreover, we discuss, relative to current practices, how and why the new inference procedures provide information about mean interrater agreement in the population, which can improve data aggregation decisions and interpretations of findings from group-level studies.

Keywords: Interrater agreement, pseudo agreement, r_{WG}, average deviation, AD, ICC

Central Tendency and Matched Difference Approaches for Assessing Interrater Agreement

Over the past several decades there has been increased interest in studying multilevel phenomena, in part thanks to important theoretical and methodological advances, including seminal works by Chan (1998) and Kozlowski and Klein (2000). Often these multilevel phenomena are theoretically grounded in consensus composition models (Chan, 1998) and justified empirically via interrater agreement statistics.[1] For instance, Myer, Thoroughgood, and Mohammed (2016) define ethical climate as organizational members' "shared perceptions regarding organizational policies, practices, and procedures that emphasize ethical content" (p. 1179). By definition, group-level (e.g., organizational) climate exists to the extent that members agree in their perceptions. For a sample of groups (e.g., a sample of organizations), researchers evaluate within-group agreement for each group, typically with r_{WG} or AD indices (LeBreton & Senter, 2007; Smith-Crowe, Burke, Kouchaki, & Signal, 2013; Woehr, Loignon, Schmidt, Loughry, & Ohland, 2015b). When summarizing within-group agreement across a sample of groups, researchers have focused on a variety of metrics: measures of central tendency (the mean or median within-group agreement across groups; e.g., Jiang, Chuang, & Chiao, 2015); the percentage of within-group agreement values above particular cutoff values (e.g., Borucki & Burke, 1999); and the range of within-group agreement values (e.g., Wallace & Chen, 2006). These summary assessments, often in conjunction with arbitrary agreement cutoff values, inform decisions concerning data aggregation. In addition, ICC(1) is commonly employed as a supplement to within-group agreement assessments and indicates the degree to which the value for any member of the group can serve as a reliable estimate of the aggregated variable (Bliese, 1998).

[1] We use the terms "within-group agreement," "interrater agreement," and "homogeneity" interchangeably. We use the term "group" to refer to any level of analysis above the individual level.

Despite their widespread use, practices around interpreting these statistics are limited. The mean or median within-group agreement across a sample of groups provides only partial information about the quality of aggregated data in these groups, as do the percentage of groups with agreement values above a particular cutoff and the range of agreement values. In particular, these descriptive statistics do not indicate whether agreement across the groups is sufficiently different from chance agreement so as to allow for the conclusion that some agreement exists regardless of the magnitude of the mean or median (cf. Cohen, Doveh, & Nahum-Shani, 2009). To date, in drawing inferences concerning whether agreement is different from “no agreement,” researchers have largely focused on single items and scales for a single group (Dunlap, Burke, & Smith-Crowe, 2003; Smith-Crowe, Burke, Cohen & Doveh, 2014) or compared the homogeneity for two or more independent groups (Cohen, Doveh, & Eick, 2001; Pasisz & Hurtz, 2009). Limited research attention has been devoted to drawing inferences based on the mean or median of a sample of groups. An exception is the work of Cohen et al. (2009) who discussed a method for obtaining critical values corresponding to a .05 significance level for mean and median r_{WG} and AD indices, and applied these methods to a sample of 49 U.S. Army companies. Yet, as pointed out by Woehr et al. (2015b), rejecting the null hypothesis of no agreement is not proof that agreement is sufficiently large to justify aggregation.

Given the limitations of the current practices, the purpose of this paper is to develop and test new ways of assessing the extent of agreement based on a sample of groups. We focus our efforts on developing methods for constructing confidence intervals around mean within-group agreement because confidence intervals indicate the range within which the population parameter (the population mean) is likely to fall, as well as precision of this estimate (i.e., the size of the range). For these reasons, the American Psychological Association (2010, p. 34) refers to confidence intervals as being “in general, the best reporting strategy” and encourages researchers to report confidence intervals whenever possible. Currently, there are no methods available for constructing confidence intervals for interrater agreement statistics.

We develop our methods using bootstrapping procedures, as well as random group resampling. We test these methods using simulated and real data. We contribute to the multilevel literature, first, by providing empirically tested ways of assessing agreement based on a sample of groups. The methods for constructing confidence intervals we present are sufficiently flexible to be useable by researchers studying any multilevel phenomena grounded in the consensus model of agreement and entailing more than one group. Second, devising these methods for constructing confidence intervals necessitated contributions that are of a more general nature and could be widely applied in multilevel research. In particular, we developed a model for multilevel data (see the Appendix), which is helpful for understanding and generating hierarchical Likert scale data. Importantly, as detailed below, we advance research relying on random group resampling for assessing overall agreement based on a sample of groups by responding to some of the problems pointed out by Woehr et al. (2015b).

In what follows, we begin by briefly reviewing key interrater agreement statistics as they are central to our measure of mean interrater agreement. We then discuss the literature on random group resampling. Whereas interrater agreement statistics are meant to be applied to single groups, and yet are applied to samples of groups, random group resampling provides an alternative way to think about agreement that is appropriately applied to samples of groups. Recent applications of random group resampling have led to suggested best practices in employing interrater agreement statistics that are unlikely to generalize to some multilevel data structures. Importantly, we illustrate how our new procedure involving random group resampling can overcome the existing limitations.

Interrater Agreement Statistics

In their recent review, Woehr, Loignon, and Schmidt (2015a) identified r_{WG} (James, Demaree, & Wolf, 1984) as the most commonly used interrater agreement statistic. For a single item, j, it is defined as

$$r_{WG(j)} = 1 - \frac{s^2}{\sigma^2}, \tag{1}$$

where s^2 is the variance in individuals' responses to item j (the observed variance) and σ^2 is the variance of the null distribution. For a scale of J items it is defined as

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

where $\bar{s}^2$ is the mean observed variance in the ratings of J items. The null distribution conceptually represents no agreement, which means that to calculate r_{WG}, one makes a direct comparison between the observed variance in individuals' ratings with the variance one would expect if there was no agreement among individuals. Higher numbers indicate greater agreement. Often researchers define no agreement, or the null distribution, in terms of a uniform distribution, where every value on a Likert scale is equally likely, as would be the case when individuals' perceptions of things are "all over the board." Yet others have written extensively about the overuse of the uniform distribution as the null distribution, pointing out that this practice results in inflated values of r_{WG} and suggesting other ways of defining the null distribution (Cohen et al., 2009; LeBreton & Senter, 2008; Smith-Crowe et al., 2013; Smith-Crowe et al., 2014). Applying a rule-of-thumb, researchers consider values of .70 and greater as indicating agreement. Increasingly, guidelines have emerged for assessing the statistical significance of a single r_{WG} value (either for an item or a scale) given a variety of null distributions (Cohen et al., 2009; Smith-Crowe et al., 2014).

The average deviation (AD; Burke, Finkelstein, & Dusig, 1999) is an alternative interrater agreement statistic that is increasingly used. It is the absolute mean difference between each individual's response and the mean response of all individuals:[2]

$$\mathrm{AD}_{\mathrm{M(j)}} = \frac{\sum_{i=1}^{N} \left| x_{ji} - \bar{x}_j \right|}{N}, \tag{3}$$

where N is the number of individuals, x_{ji} is the i'th individual's response to item j, and $\bar{x}_j$ is the mean response to item j. AD_{M} for scales ($\mathrm{AD}_{\mathrm{M(J)}}$) is calculated as the mean of the $\mathrm{AD}_{\mathrm{M(j)}}$ values for J items. AD assesses the degree of dispersion in individuals' responses, which means that lower values indicate greater levels of agreement. Unlike r_{WG}, for which the calculation of the statistic entails a comparison of observed variance to the variance of a null distribution, for AD the comparison to a null distribution comes at the stage of interpreting the calculated AD value. That is, interpreting AD values is based on a judgment of what would be a sufficiently low observed value relative to the dispersion associated with the theoretical null distribution representing no agreement. There are extensive guidelines available for assessing the significance of AD values relative to a variety of null distributions (Burke & Dunlap, 2002; Cohen et al., 2009; Smith-Crowe et al., 2013; Smith-Crowe et al., 2014).

In addition to reporting interrater agreement statistics, multilevel researchers also commonly report ICC(1) values, which indicate the extent to which individuals' responses are attributable to group membership (LeBreton & Senter, 2008). Essentially ICC(1) combines the

[2] The average deviation can also be calculated relative to the median ($\mathrm{AD}_{\mathrm{Md}}$). Here we focus on AD_{M} in the interest of space.

notions of agreement and reliability (LeBreton & Senter, 2008). Based on a one-way random effects ANOVA, ICC(1) is defined as follows:

$$\mathrm{ICC(1)} = \frac{\mathrm{MS_B} - \mathrm{MS_W}}{\mathrm{MS_B} + (\mathrm{n}-1)\mathrm{MS_W}}, \tag{4}$$

where MS_B is the mean square between groups, MS_W is the mean square within groups and n is the group size (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Thus, ICC(1) compares within group variance to between group variance, with higher values indicating increasingly low variability within groups and high variability across groups. LeBreton and Senter have recommended interpreting ICC(1) values as effect sizes, where small=.01, medium=.10, and large=.25. Yet, interpreting these values is not necessarily straightforward.

LeBreton, Burgess, Kaiser, Atchley, and James (2003) demonstrated that reliability and agreement can be negatively correlated. Notably, if there is little difference between groups, ICC(1) values will be low, possibly despite high agreement within groups. Such values would be misleading in the context of consensus models, which specify within group agreement as the standard, not variance between groups. LeBreton et al. concluded that to employ both agreement indices (e.g., $\mathrm{r_{WG}}$ and AD) and hybrid indices (e.g., ICC which assesses agreement and reliability) is to employ "psychometric checks and balances" (p. 121). In this case, we suggest that one would not want to presume that aggregation is not justified based solely on low ICC(1) values. We note that, with some exceptions (e.g., see Gil, Rico, Alcover, & Barrasa, 2005; Neal, West, & Patterson, 2005; Riordan, Vanderberg, & Richardson, 2005), researchers very often justify data aggregation decisions with respect to both an interrater agreement statistic and ICC(1) (e.g., see Chen, Liu, & Portnoy, 2012; Fahr, Lee & Fahr, 2010 Gonzalez-Roman & Hernandez, 2014; Gupta et al.,2018; Jiang, Chuang, & Chiao, 2015; Morrison, Wheeler-Smith, & Kamdar, 2011; Naveh & Katz-Navon, 2015; Neal & Griffin, 2006).

A limitation of the existing guidelines for assessing agreement is that they largely focus on inferring the level of agreement within a single group, rather than agreement in the population of groups. We argue that estimating interrater agreement as a population parameter and establishing a confidence interval for the population parameter are justified and important goals. First, in almost all studies in the literature using interrater agreement indices to support data aggregation, there is a sample of groups from a larger set of groups or population (e.g., see Bain, Mann, & Pirola-Merlo, 2011; Chuang & Liao, 2010; Jiang et al., 2015; Menges, Walter, Vogel, & Bruch, 2011; Naveh & Katz-Navon, 2015). For instance, Jiang et al. randomly sampled 200 out of 400 shoe stores (with data being supplied by 142 stores), where the author used the mean interrater agreement value to justify data aggregation. Consistent with Jiang et al., most researchers rely exclusively on summaries of the overall sample, including means, medians, and percentages of interrater agreement statistics exceeding a threshold to justify data aggregation (e.g., see Bass, Avolio, Jung, & Berson, 2003; Bunderson, 2003; Kirkman, Tesluk, & Rosen, 2001; Sowinski, Fortmann, & Lezotte, 2008; Takeuchi, Chen & Lepak, 2009).

Second, in virtually every case where a sample of groups is obtained, the author makes general conclusions or generalizes the findings to other groups, organizations, or contexts (e.g., see Berson, Da'as, & Waldman, 2015; Cole, Carter, & Zhang, 2013; Gonzalez-Roma & Hernandez, 2014; Wallace & Chen, 2006). For example, Wallace and Chen (2006), based on a sample of 50 workgroups from a facilities department of a large university, made conclusions as to how their findings related to "work organizations" in general. Notably, in some cases, the author generalizes the interrater agreement findings themselves. An example of the latter practice is where Simons and Roberson (2003) concluded "First, this study builds on Mossholder

(1998) by adding empirical support for the validity of aggregate justice perceptions in organizations as a focus for research" (p. 442).

The fact that authors are sampling groups, using mean or median interrater agreement values to justify aggregating data for the sample of groups, and then making general group-level conclusions calls for a better understanding of interrater agreement as a population parameter. Basically, when we assess agreement and use the mean or median to represent agreement in a sample of groups, we are also inferring what agreement would be observed in different groups (or organizations) had we sampled those groups (or organizations). Estimating interrater agreement as a population parameter and establishing a confidence interval around the population parameter will show researchers the range of plausible values for mean interrater agreement in the population. As we illustrate below, knowledge of this plausible range within a primary group-level study can better inform data aggregation decisions and interpretations of group-level findings insofar as their potential generalizability is concerned.

Third, in almost all studies where individual level data are aggregated, the author does not eliminate any group on the basis of a low or out-of-bound interrater agreement result, and the author decides to aggregate even when mean (or median) interrater agreement values are below rules-of-thumb cutoff values (e.g., see Binci, 2011; Boehm, Dwertmann, Kunze, Michaelis, Parks, & McDonald, 2014; Borucki & Burke, 1999; de Jong, de Ruyter., & Lemmink, 2005; Hofmann & Mark, 2006; McKay, Avery, & Morris, 2009; Morrison et al., 2011; Patterson et al., 2002; Simons & Roberson, 2003). In a very small number of studies, the researcher will eliminate groups based on r_{WG} values being below .70 (e.g., see Aryee, Chen, & Budhwar, 2004; Riordan et al., 2005; Susskind, Kacmar, & Borchgrevink, 2003). These practices indicate that the vast majority of authors do not consider interrater agreement to be a characteristic of each group in their investigation, and that acceptable interrater agreement in each group is not considered as a necessary condition for data aggregation. The general practice of retaining all groups from a sample despite the fact that some groups have low interrater agreement is another reason for obtaining a better understanding of interrater agreement as a population parameter.

Random Group Resampling (RGR)

Bliese and Halverson (1996, 1998, 2002) introduced an alternative approach to assessing agreement, which does examine overall agreement in a sample of groups: random group resampling (RGR). The logic of this procedure is that true group effects can be identified when the results of real groups are significantly different from the results of pseudo groups. Pseudo groups are formed by randomly combining individual responses into groups (Bliese & Halverson, 2002; Woehr, Loignon, & Schmidt, 2015a). Bliese and Halverson (2002) demonstrated the use of this procedure using multilevel data from 2,042 individuals within 49 army companies. Their hypothesis was that a supportive leadership climate within a company would mitigate the relationship between task significance (a stressor at the company level) and hostility (a strain at the company level). From their original data, they created 1,000 samples of pseudo groups, each with 49 groups composed of a total of 2,042 individuals randomly assigned to groups without replacement; the group sizes in the samples of pseudo groups equaled the group sizes of the real groups, which varied across groups. They compared the hierarchical regression results from the real groups (companies) against the samples of pseudo groups. For the real groups, there were significant main effects for task significance and leadership climate, as well as a significant interaction effect. Further, they found that the mean squares for task significance, leadership climate, and the interaction were significantly greater in the real groups compared to the pseudo groups, and that in each case the mean squares in the real groups exceeded the upper bounds of the 95% confidence intervals created from the samples of pseudo groups. Notably, they also found significant main effects for task significance and leadership climate, but not for the interaction, in the pseudo groups. Their conclusion was that there is a true group effect, though some portion of the significant main effects found for the real groups can be attributed to pseudo agreement.

Other researchers have used RGR to study the functioning of interrater agreement statistics (Cohen et al., 2009; Ludtke & Robitzsch, 2009). In a recent large-scale investigation employing RGR with real and pseudo groups based on CATME (Comprehensive Assessment of Team Member Effectiveness) data, Woehr et al. (2015b) examined the extent to which r_{WG}, AD, and ICC indices reflect real agreement versus "pseudo-agreement.” The latter is indicated by similar levels of agreement in real groups compared to pseudo groups, the logic being that a group effect is indicated when real groups have greater agreement than pseudo groups. Among the indices studied Woehr et al. showed that unlike AD and $r_{WG,}$ ICC(1) identified the difference between real and pseudo groups. They suggested that researchers should interpret ICC values as an initial hurdle for determinations of agreement for a sample of groups, where agreement is indicated by ICC values for real groups sufficiently exceeding those for pseudo groups. We note, however, that their data were highly skewed and restricted in their range, which may limit the generalizability of their proposed hurdle approach. We also note that the proposed procedure does not allow for assessments of whether differences between real and pseudo groups are due to chance. Below, we expand on the work of Woehr et al. by introducing a novel inference method that uses RGR and enables one to make conclusions about agreement vs. pseudo-agreement.

Advancing Current Practices

Our review of agreement related practices within the multilevel research literature highlights two problems. First, interrater agreement statistics for assessing single groups are applied to samples of groups in a way that provides only limited information about agreement in the population of groups. We address this limitation in Study 1 by developing a “central tendency approach” via simulated multilevel data and bootstrapping for constructing confidence intervals around population parameters of r_{WG}, AD, and ICC such that researchers, with 95% confidence, can infer that the true level of agreement within the population falls within a particular range.

Second, the latest guidelines for the alternative random group resampling (RGR) approach, which does entail assessing agreement across groups, may not generalize broadly and do not allow for calculations of the likelihood that differences between real and pseudo groups are due to chance. We address this limitation in Study 2 by developing a new approach, using RGR, which considers a matched difference between assessments of the respective r_{WG}, AD, and ICC values for real and pseudo groups, with the establishment of confidence intervals around such differences. This new, “matched difference approach,” and procedures for the respective interrater agreement indices are intended to better inform researchers as to whether the difference between real and pseudo agreement across the groups is sufficiently different from chance to conclude that the homogeneity within groups is real. We used actual team data, as well as simulated data to evaluate the efficacy of our new inference procedures for differentiating real and pseudo agreement.

Study 1

The overarching purpose of Study 1 was to define population parameters for overall agreement based on a sample of groups with respect to r_{WG} and AD, and to use data generated from an underlying model, where individuals are nested within groups, to construct confidence intervals around the population parameters. We also establish confidence intervals for ICC(1) given that it is almost always included with an interrater agreement statistic to justify data aggregation. In establishing confidence intervals for r_{WG}, AD, and ICC(1), another aim of Study 1 was to show how confidence intervals add information to summary statistics for observed agreement within groups, providing information not found in the summary statistics alone but information that is useful for assessing the extent of agreement across groups and interpreting the potential generalizability of study findings.

Method

In constructing confidence intervals around mean interrater agreement statistics, to be defined below, we use both simulated and real data. The value of using simulated data is that with defined parameters, we are able to draw conclusions about accuracy. Generating multivariate hierarchical data with a Likert scale, however, posed a considerable challenge due to the complexity of simulating such data and the limitations of current methods. In previous research on agreement indices, data were simulated for single groups (Smith-Crowe et al., 2014), not for individuals nested in groups. We generalized earlier procedures in Smith-Crowe et al. (2014) by developing an underlying model upon which we based our data generation method. The model and the data generating process based on the model uniquely take into account the intra-class correlation due to similarity within groups as well as the correlation between the items within subjects. In the Appendix, we report and discuss our model, which can be applied broadly to multilevel investigations using simulated data. In the Supplemental Materials, we show how the parameters in the model relate to r_{WG}, AD, and ICC(1), which provides an understanding of the circumstances under which one will observe various magnitudes of within-group agreement and ICC values.

We generated two simulated datasets, which we designate as “SIM1A” and “SIM1B.” The simulated data were generated so as to relate to typical conditions in a group-level study. While the characteristics of group-level studies vary, there are numerous examples in the literature where the number of individuals per group is between 5 and 10 (e.g., Bain, Mann, & Pirola-Merlo, 2001; Chuang & Liao, 2010; Smith, Collins, & Clark, 2005), the number of groups is between 25 and 50 (e.g., Brahm & Kunze, 2012; de Jong et al.,, 2005; Probst, 2015), the number of items per measure is between 3 and 9 (Boehm, Kunze, & Bruch, 2014; Borucki & Burke, 1999; Towler, Lezotte, & Burke, 2011), and the number of response options per item is 5 or 7 (e.g., Dawson, Gonzalez-Roma, Davis, & West, 2008; Dong, Liao, Chuang, Zhou, & Campbell, 2015; Liu, Hernandez & Wang, 2014). Given these common features of group-level studies, we set the following parameters for both SIM1A and SIM1B: the number of items (J) equaled 6; the number of Likert scale response options (A) equaled 5; the number of individuals per group (n_k) equaled 5; and the number of groups (K) equaled 30.

The model parameters[3] for both datasets were as follows: β=0.7, σ_ε=0.45, σ_δ=1, ω_1=30, ω_2=4, θ_1=0.8, and θ_2=0. The underlying model was the uniform distribution, such that each of the A values is equally probable (1/5). While the underlying model is the uniform distribution, introducing a group effect (symbolized as α) will cause the entire group distribution of responses to shift on the Likert scale. For SIM1A, we selected a smaller group effect, expressed by a smaller variation for alpha[4] (σ_α=0.6), causing a slight shift in the overall group distribution of responses and relatively low agreement. In SIM1B, we selected a larger group effect (σ_α=2.5), causing a large shift in the group distribution of responses and higher agreement. As described

[3] See the Appendix for an explanation of the model parameters, and the Supplemental Materials for a more detailed explanation of omega.

[4] See the Appendix for a more detailed explanation of alpha.

in more detail in the Appendix, the group effect expressed by alpha is the shift in the entire distribution of responses of the group, while the introduction of a group shrinkage effect, refers to whether the range of the distribution of responses within the group increases or decreases. For the purposes of Study 1, we did not introduce a group shrinkage effect, only group effects were introduced in SIM1A and SIM1B; we did employ group shrinkage effects in Study 2.

Also, we note that $r_{WG(J)}$ was calculated using the uniform distribution as this null distribution. With only a few exceptions (e.g., Cole, Carter, & Zhang, 2013; Eisenbeiss, Knippenberg, & Boerner, 2008), the vast majority of authors have used a uniform response distribution when computing r_{WG} (e.g., see Gonzalez-Roma & Hernandez, 2014; Griffith, 2006; Hofmann & Mark, 2006; Katz-Navon, Naveh, & Stern, 2005; MacCormick & Parker, 2010; Probst, 2015; Wallace & Chen, 2006; Wallace, Johnson, Mathe, & Paul, 2011). Furthermore, there is insufficient theory to specify an alternative null response distribution, such as a particular skewed distribution, that would apply across all group-level studies. Together, these considerations led us to use the uniform distribution for the purposes of our simulation work.

We used these data to construct confidence intervals around unknown parameters representing mean group agreement of the population of groups. For this "central tendency approach" and procedures, we defined population parameter estimates as

$$\hat{\rho}_{\mathrm{WG(J)}} = \frac{\sum_{k=1}^{K} r_{\mathrm{WG(J)}}(k)}{K}, \tag{5}$$

where $r_{WG(J)}(k)$ denotes the r_{WG} value obtained for group k, and

$$\widehat{\mathrm{AD}}_{\mathrm{M(J)}} = \frac{\sum_{k=1}^{K} \mathrm{AD}_{\mathrm{M(J)}}(k)}{K}, \tag{6}$$

where $AD_{M(J)}(k)$ denotes the $AD_{M(J)}$ value obtained for group k. We also used these data to construct confidence intervals around the unknown parameter for ICC(1). We assume that the sample of groups (K) is a random sample from the population of groups. Because the distribution of these statistics cannot be expressed in closed form, we used bootstrapping techniques for the construction of the confidence intervals. When the number of groups is relatively large and the group sizes are small (as true for SIM1A and SIM1B), the *B* bootstrap samples should be generated by resampling only the groups (see Davison & Hinkley, 1997; Ren, Lai, Tong, Aminzadeh, Hou, & Lai, 2010). Here, *B* denotes the number of bootstrap samples. When the group sizes differ (unlike SIM1A and SIM1B), the total bootstrap sample may vary slightly between the *B* bootstrap samples. We assessed how accurate these resampling schemes are in producing a confidence interval that included the known parameters when group sizes differ. As discussed in more detail below relative to the simulation conditions for Study 2, our findings confirm what is known concerning resampling only groups when group sizes are small. In the Supplemental Materials, we present results for the two resampling schemes.

For both SIM1A and SIM1B Monte Carlo runs, 1000 samples were generated to obtain the "true" parameters corresponding to the statistics $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1). In doing so, we assume the median value from each of the 1000 samples to be the “true” population value. Then, for both SIM1A and SIM1B, we randomly selected one of the 1000 Monte Carlo samples (as we do practically in research using real data). For each one of these two samples, we constructed the bootstrap confidence intervals for each of the three parameters. The parameter estimates were calculated using Equations 4-6, and were based on the 30 groups in each sample.

In addition to employing simulated data, we also constructed confidence intervals using archival data. These data were collected as part of a survey called the Technion Multicultural Team Project (TMCTP), the purpose of which is to provide MBA students in courses on cross-cultural and global management with hands-on experiences in virtual, multicultural teams. The TMCTP is a collaboration involving business schools around the world. The students work in teams of three to four members for approximately three weeks to create a business proposal. The data we used are part of a larger dataset of 1,221 MBA students collected in four time points (see Erez et al., 2013); here we are working in each study with data from only one of the four times. The data we used in the current study, which we designate as "TMCTP," includes 60 teams, the majority of which have four members, for a total of 229 individuals. We analyzed individuals' responses to two 3-item measures: task conflict employing a 7-point Likert scale (Mannix, 2001), and team identity employing a 5-point Likert scale (Earley & Mosakowski, 2000).

Results and Discussion

The results for SIM1A, SIM1B, and the TMCTP datasets are reported in Table 1. For each dataset, the observed agreement and ICC(1) statistics corresponding to the 2.5 and 97.5 percentiles are reported, as well as the median values (50 percentile). As noted above, we consider the median value of the Monte Carlo samples to be the "true" population parameters for $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1), respectively. The parameter estimates and 95% confidence intervals are based on the single samples randomly selected from SIM1A and SIM1B (i.e., we used one sample from each of these datasets). As we can see in Table 1, the confidence intervals contain the population parameters in all cases.

Insert Table 1 about here.

For SIM1A, the parameter estimate for the single sample for $r_{WG(J)}$ was 0.596 with a 95% confidence interval of 0.467 to 0.713. Typically, when an average $r_{WG(J)}$ of approximately 0.60 is obtained, the author will refer to James (1982) or Glick (1985) and a cut-off value of 0.60 to justify data aggregation. Here, the author often indicates that the observed $r_{WG(J)}$ "approximates the lower-bound rule-of-thumb" or is slightly greater than the "minimal acceptable value of 0.60

or 0.60 cutoff" (e.g., Binci, 2011; Hui, Chiu, Yu, Cheng, & Tse, 2007; McKay, Avery, & Morris, 2009). And in some cases, the author will justify data aggregation by reporting the percentage of $r_{WG(J)}$ values above 0.60 (e.g., see Borucki & Burke, 1999). For the 30 groups in the SIM1A sample, 70% of the $r_{WG(J)}$ values were at or above 0.60. Authors such as Borucki and Burke (1999) considered such a percentage as indicating acceptable interrater agreement.

Notably, the lower limit of the 95% confidence interval for SIM1A was 0.467, with 56% of the confidence interval being below the 0.60 value. The lower limit of 0.467 as well as the large percentage of plausible mean interrater agreement values below 0.60 indicate that mean interrater agreement may not be acceptable, in a practical sense, in the population. In addition, the width of the interval (0.47 to 0.71), or margin of error in estimating mean $r_{WG(J)}$, is large. As such, the confidence interval itself would suggest that the researcher use caution when deciding to aggregate such data, and when interpreting group-level relationships and their potential generalizability based on such data.

Also for the sample of groups in SIM1A, the parameter estimate for $AD_{M(J)}$ of 0.868 is above the practical cutoff of 0.85 for a 5-point Likert type scale and the uniform null response distribution (see Smith-Crowe et al., 2013). For the 30 groups in the SIM1A sample, 60% of the observed $AD_{M(J)}$ values were above the cutoff value. Typically, the researcher moves ahead with data aggregation with large percentages of unacceptable AD values (e.g., see McKay et al., 2009). For the 30 groups in the SIM1A example, a likely interpretation of these findings would be that the mean $AD_{M(J)}$ is only slightly above the cutoff value, but may be considered close enough to justify data aggregation.

However, the upper limit of the 95% confidence interval for $AD_{M(J)}$ in SIM1A was 0.947, with 59% of the confidence interval being above the 0.85 cutoff value. The upper limit of

0.947, as well as the large percentage of plausible mean interrater agreement values above 0.85, indicate that mean interrater agreement as estimated by the $AD_{M(J)}$ index may not be acceptable, in a practical sense, in the population. In addition, the width of the interval (0.78 to 0.95), or margin of error in estimating mean $AD_{M(J)}$, is large. Consistent with conclusions about $r_{WG(J)}$ for SIM1A, the confidence interval itself for $AD_{M(J)}$ would suggest that the researcher use caution when deciding to aggregate such data and interpreting group-level relationships based on such data.

On the other hand, the ICC(1) parameter estimate of 0.402 for SIM1A is above typical rule-of-thumb minimum values for ICC(1) of 0.05 to 0.12 (Bliese, 2000; James, 1982), and the 95% confidence interval for ICC(1) is above 0.12. While the latter confidence interval finding points to a group effect, the overall pattern of findings relative to confidence intervals for $r_{WG(J)}$ and $AD_{M(J)}$ would raise questions as to whether mean interrater agreement is acceptable, in a practical sense, in the population. This conclusion and cautionary note about data aggregation would not have been reached in the typical group-level study that relies exclusively on point estimates (i.e., mean or median values) for interrater agreement and practical cutoff values for acceptable agreement.

For the single sample from the SIM1B dataset and for the two scales in the TMCTP data, the respective parameter estimates for $r_{WG(J)}$ and 95% confidence intervals were 0.877 (0.800, 0.937), 0.809 (0.761, 0.852), and 0.949 (0.936, 0.959) respectively. Typically, when an average $r_{WG(J)}$ of approximately 0.81 is obtained, the author will refer to one or more of a large number of sources (e.g., Bliese, 2000; George, 1990; James, Demaree, & Wolf, 1993; Kozlowski & Klein, 2000) and a value of 0.70 to justify aggregation. Here, the author frequently indicates that the average $r_{WG(J)}$ is above 0.70, suggesting a “good amount of agreement” or “sufficient agreement” (e.g., see Abdelhadi & Drach-Zahavy, 2011; Brahm & Kunze, 2012; Koene, Vogelaar, & Soeters, 2002). Notably, the lower limits of the 95% confidence intervals for the SIM1B and TMCTP datasets for r_{WG} were above 0.70. These findings indicate that all plausible values for mean r_{WG} are acceptable with respect to the 0.70 rule-of-thumb and other standards for strong agreement (see LeBreton & Senter, 2008). Furthermore, the widths of the confidence intervals are generally narrow, indicating that the margin of error in estimating mean $r_{WG(J)}$ is small. In this regard, knowledge of the confidence interval for $r_{WG(J)}$ provides additional information relative to a point estimate for mean interrater agreement and for a more informed decision concerning data aggregation.

The parameter estimates for $AD_{M(J)}$ for the single sample from the SIM1B and the two measures from the TMCTP dataset of 0.421, 0.824, and 0.485 are below the practical cutoffs for $AD_{M(J)}$ for 5-point (i.e., 0.85) and 7-point (i.e., 1.21) Likert-type scales, respectively (Smith-Crowe et al., 2013). The 95% confidence interval for SIM1B is (0.290, 0.551) and for the two TMCTP measures are (0.755, 0.899) and (0.441, 0.535). Notably, these confidence intervals for $AD_{M(J)}$ do not include the upper limit cutoffs for 5-point and 7-point scales. As such, these findings indicate that all plausible values for mean AD are acceptable with respect to the practical cutoffs for this index (see Smith-Crowe et al., 2013). A strong group effect in the population is also indicated in SIM1B and in one of the two TMCTP data sets in that the 95% confidence intervals for ICC(1) lie well above typical rule-of-thumb minimum values for ICC(1) of 0.05 to 0.12 (see Bliese, 2000, James, 1982). For Task Conflict, the lower bound is slightly below the stricter rule. Together, knowledge of the range of plausible values for the $AD_{M(J),}$ and ICC(1) parameters would provide useful, additional (relative to just the sample mean or sample

ICC(1) value) information in making a decision to aggregate data from the SIM1B and TMCTP datasets and assist in better interpreting group-level findings involving such variables.

In sum, the above examples illustrate the additional information obtained by using confidence intervals for assessments of overall agreement and the potential usefulness of such information relative to data aggregation decisions in group-level studies and interpreting findings from primary group-level studies. In Study 2, we extend the use of confidence intervals to random group resampling (RGR) to deal with the problem of pseudo agreement and the question of whether differences between pseudo and real agreement are due to chance.

Study 2

The purpose of Study 2 is to present a new inference approach and procedures for distinguishing between real and pseudo agreement and to provide guidance on the construction of confidence intervals for inferring that homogeneity within groups is real. Here “real” agreement is indicated when the homogeneity of actual groups is sufficiently greater than that of reshuffled, pseudo groups. We expand on recent research by Woehr et al. (2015b), who examined the distributional characteristics of r_{WG}, AD, and ICC values. Using random group resampling (RGR; Bliese & Halverson, 1996), they found that ICC values distinguished between pseudo and real agreement, but that $r_{WG(J)}$ and AD did not. Given the skew and range restriction of their data, we assume that their results might have limited generalizability. We expand on their work by presenting a “matched difference approach” and procedures that are more likely to distinguish between pseudo and real agreement across a variety of circumstances.

Method

Our procedure was as follows. For a given dataset, we generated B=1000 bootstrap samples. Then, for each bootstrap sample (with K groups and n_k individuals within each group), we performed random group resampling (RGR). From these results, we recorded the resulting pairs of $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1) values obtained from the bootstrap sample and its matched, reshuffled sample, and we calculated the difference between the pairs (e.g., the difference between the $r_{WG(J)}$ value for a given bootstrap sample and the $r_{WG(J)}$ value for its matched reshuffled sample). From the differences obtained from each of the B samples, we constructed a confidence interval around their mean. The bootstrap (1 - p) confidence intervals were constructed by using the percentiles of the bootstrap distribution (the p/2 and 1 - p/2 percentiles). When there is no difference between the statistics observed for a sample and for its reshuffled matched sample, then we infer that any homogeneity indicated in the sample is not real. Thus, if zero is included within the confidence interval, it indicates pseudo agreement is plausible in the population, and if the confidence interval does not include zero, it indicates that some homogeneity within groups is real or plausible in the population. Notably, by using differences between matched samples of shuffled vs. not shuffled samples, we gain power similar to the gain obtained by using matched t-tests to compare means of matched samples.

To illustrate our method, as in Study 1, we used the task conflict data (a three-item measure with a 7-point Likert response scale) and the team identity data (a three-item measure with a 5-point Likert response scale) from the TMCTP dataset. The "true" answer of whether there is group consensus sufficient to justify aggregation in the TMCTP data is unknown. Thus, in using the TMCTP data alone, we cannot assess whether the method that we introduce provides accurate answers.

In order to evaluate the accuracy of our method, we also include simulated data examples, for which we know the underlying model. We generated five datasets, which we designate as “SIM2A,” “SIM2B,” “SIM2C,” “SIM2D,” and “SIM2E.” These simulated data complement our use of real data. The simulated data were generated using the same method described in Study 1 (see also the Appendix). As was true for Study 1, the uniform distribution was the underlying distribution for all of the simulated datasets. Further, these parameters were common to all datasets: the number of items (*J*) equaled 6; the number of Likert scale response options (*A*) equaled 5; the number of individuals per group (n_k) equaled 10; and the number of groups (*K*) equaled 50. Further, β=0.7, σ_ε=0.2, σ_δ=1, ω_1=1, and ω_2=1.

The parameters specific to each dataset are shown in Table 2. As shown in this table, we varied the group effect or variance between groups (denoted by the standard deviation of alpha, σ_α) to reflect no group effect, as well as to include group effects ranging from smaller to larger. In addition, we varied the degree to which the range of the response distribution contracted, expanded, or remained the same, referred to in Table 2 as shrinkage, anti-shrinkage, and no shrinkage, respectively. The combinations of group and shrinkage effects allowed us to assess the accuracy of our proposed procedure under a variety of conditions including a less likely condition involving no group effect and a more extreme condition concerning the expansion of the response distribution with a group effect.

Insert Table 2 about here.

Results and Discussion

We begin by reporting the results for the TMCTP data. Since the number of groups was relatively large (*K*=60) and the size of the groups was small (n_k=3 to 4), the *B* bootstrap samples were generated by resampling only groups and not individuals. Thus, each bootstrap sample included 60 groups as was the structure of the original sample. However, since the group sizes were not equal, the total number of individuals in the bootstrap sample varied slightly between the bootstrap samples. Here, for the task conflict data, the time point we chose differed from the one in Study 1, since its structure was of more interest. The results for the TMCTP dataset are shown in Table 3. The confidence intervals for the difference between the bootstrap samples and their matched, shuffled sample are bolded because it is this interval that is the focus of our proposed inference procedure.

Insert Table 3 about here.

For the sampled data for task conflict, $r_{WG(J)}$ =0.824, $AD_{M(J}$=0.792, and ICC(1)=0.0. Recall that according to Smith-Crowe et al. (2013), AD values should be smaller than 1.21 for a 7-point scale in order to conclude that practically significant agreement exists; according to the commonly applied rule-of-thumb for $r_{WG(J)}$, values should be larger than 0.70 in order to conclude that there is a meaningful degree of agreement. Based on these criteria, we conclude that there is agreement in the population. Yet, the 0.02 value for ICC(1) indicates that there is no consistency (i.e., there is no group effect). Indeed, all of the 95% confidence intervals for the differences between the bootstrap and matched, shuffled samples include zero: -0.027, 0.037 for $r_{WG(J)}$; -0.062, 0.052 for $AD_{M(J)}$; and -0.172, 0.160 for ICC(1). These results indicate that the agreement indicated by the $r_{WG(J)}$ and $AD_{M(J)}$ values for the sampled data (0.824 and 0.792, respectively) represent agreement observed among the individuals in the whole sample, and not just among individuals within groups. That is, here we are seeing pseudo rather than real agreement. Importantly, however, while the designation of “pseudo” appears as if it would be universally problematic given the negative connotation of the term, it is not. We discuss the nuances of interpretation in the general discussion.

For the sampled data of the team identity scale, $r_{WG(J)}$=0.949, $AD_{M(J)}$=0.485, and ICC(1) =0.244. As for the task conflict measure, here too, in the original data, $r_{WG(J)}$ is high and $AD_{M(J)}$ is low, indicating agreement. However, unlike the previous example, here we see a "group" effect according to the magnitude of the ICC(1) value, which is greater than the rule-of-thumb thresholds of .05 to .12 (Bliese, 2000; James, 1982). The confidence intervals reported in Table 3 support this conclusion as well because the ICC(1) bootstrap confidence interval does not include zero, 0.7 is below the confidence interval of $r_{WG(J)}$, and 1.21 is above the confidence interval for $AD_{M(J)}$. When the data are shuffled, the values for 95% confidence interval for ICC(1) are considerably lower than that for the bootstrap sample, such that the upper limit for the shuffled sample is less than the lower limit for the bootstrap sample. Similar to the example given by Woehr et al. (2015b), the confidence intervals for the shuffled and not shuffled agreement indices ($r_{WG(J)}$ and $AD_{M(J)}$) are much more overlapping. However, when we examine the confidence interval of the differences between the matched samples, we see that all values are positive for $r_{WG(J)}$ and negative for $AD_{M(J)}$, meaning that zero is not included in these intervals. Thus, from the matched samples it can be concluded that after the shuffling there is a significant decrease in $r_{WG(J)}$ and an increase in $AD_{M(J)}$, meaning that agreement is less apparent in the shuffled, pseudo groups.

This example supports the claim of Woehr et al. (2015b) that $r_{WG(J)}$ and $AD_{M(J)}$ unlike ICC(1) may fail to differentiate between the agreement observed in original versus the shuffled samples. However, when we use the matched differences between the paired non-shuffled and shuffled values of ICC(1), $r_{WG(J)}$, and $AD_{M(J)}$, we can assess whether there is "real" agreement in the data, beyond apparent, but nevertheless "fake" homogeneity implied by just the shuffled values of ICC(1). Moreover, we can also provide a measure of its size. That is, the median of the difference indicates that $r_{WG(J)}$ increases by .012 due to true homogeneity. It is estimated to be a 0.043 decrease according to $AD_{M(J)}$. This example demonstrates the advantage of our method, as it uses the shuffling to identify the contribution of the real groups. That is, we gain interpretative power when we examine the confidence interval of the difference between the bootstrap and matched, shuffled sample, compared to examining the unmatched samples.

Next, we present the results from our five simulated datasets. Because we know the underlying model of these datasets, these data allow us to assess the accuracy of our proposed procedure. For each dataset, we present the bootstrap results based on a single randomly selected sample out of the B=1000 samples generated. In order to learn about the stability of the procedure we repeated this procedure twice for each dataset. Since the results were essentially the same for the two repetitions, in Table 4 we present only one of the two repetitions for each of the five simulated datasets. Additionally, in the Supplemental Materials graphical displays of the results presented in Table 4 are provided.

Insert Table 4 about here.

For SIM2A, the dataset was constructed such that it entailed shrinkage (a contraction of the distribution of responses) but no group effect (i.e., no variance between groups; see Table 2). The $r_{WG(J)}$ and $AD_{M(J)}$ values for the sample indicate agreement; the $r_{WG(J)}$ was 0.769, which indicates high agreement within groups (LeBreton & Senter, 2008) and the $AD_{M(J)}$ value was .755, which is below the .85 cut-off for a 5-point Likert scale measure (Smith-Crowe et al., 2013). The value of ICC(1) was 0.051. According to Bliese (2000), ICC(1) values exceeding 0.05 are considered sufficient to warrant aggregation, which makes our observed value on the borderline of the cutoff. We would expect this borderline result because the sample was drawn from a population with no group effect. More importantly, all confidence intervals for differences include zero. This result suggests the accuracy of our procedure because the intervals correctly indicate that there is no group effect in the population.

For SIM2B, the dataset was constructed such that it entailed a group effect but no shrinkage (see Table 2). The $r_{WG(J)}$ and $AD_{M(J)}$ values for the sample both indicate substantial agreement (LeBreton & Senter, 2008; Smith-Crowe et al., 2013). The value of ICC(1) was also high, indicating a large group effect (Bliese, 2000). This result was expected since the sample was drawn from a population with a sizeable group effect. Correspondingly, none of the confidence intervals for differences include zero, and the magnitude of the differences between the bootstrap and reshuffled samples are notable. Like the results of SIM2A, these results suggest the accuracy of our inference procedure because they are aligned with the model parameters for SIM2B.

For SIM2C, the dataset was constructed such that it entailed both a group effect and shrinkage (see Table 2). The $r_{WG(J)}$ and $AD_{M(J)}$ values for the sample not only indicate substantial agreement (LeBreton & Senter, 2008; Smith-Crowe et al., 2013), but due to the addition of shrinkage in this simulation, the magnitudes of these statistics indicate greater agreement than was seen in SIM2B where there was no shrinkage. The value of ICC(1) is also high, indicating a large group effect (Bliese, 2000). This indication of a large group effect was expected because the sample was drawn from a population with both a group effect and shrinkage. The group effect expressed by large variation between groups, increases ICC(1), while the shrinkage decreases the variance within the groups. Moreover, as expected, there are no zeros in the confidence intervals of the differences between the bootstrap and matched, shuffled samples, and the magnitude of these differences is considerable.

For SIM2D, the dataset was constructed similarly to that of SIM2C such that it entailed both a group effect and shrinkage, but in the case of SIM2D the group effect was set to be smaller than in SIM2C (see Table 2). Given the similarities in populations, the results of these data are similar to that of SIM2C except that the values of $r_{WG(J)}$ and ICC(1) are lower, and the value of $AD_{M(J)}$ is higher, yet their magnitudes indicate both substantial agreement and a meaningful group effect. Again, we see no zeros in the difference confidence intervals, and notable magnitudes of differences; these results are consistent with the group effect at the population level.

For SIM2E, the dataset was constructed such that it entailed a group effect and anti-shrinkage (i.e., expansion; see Table 2). We explain the data structure by referring to the simulated data. In this example, the deltas that are producing the variances within the groups were multiplied by 2, the resulting simulated delta for each individual in the same group is rather extreme (either "large positive" or "very small negative"). Thus $r_{WG(J)}$ will be mostly low and $AD_{M(J)}$ will be mostly large. The group effect alpha shifts the whole group either to the positive or negative side. In some rarer cases, it reduces the heterogeneity among the individuals within the group resulting in agreement indices indicating greater within-group agreement. Therefore, the data includes essentially two types of groups. This example is an unusual example that we present in order to demonstrate situations when ICC(1) indicates a group effect though there is no within-group agreement. The $r_{WG(J)}$ of the sample was 0.363 and $AD_{M(J)}$ was 1.079, which indicate low within-group agreement; however, the value of ICC(1) was sufficiently high at 0.383 to indicate a group effect. Further, the group effect in the population is indicated by the confidence intervals, where there are no zeros included in any of the intervals of the difference between the bootstrap and matched samples, and the magnitudes of the differences are notable.

Given that our methods were based on resampling only a certain number of individuals per group (i.e., 5 in Study 1 and 10 in Study 2) from a certain number of groups (i.e., 30 in Study 1 and 50 in Study 2), one might ask whether our results would generalize to other group sizes and numbers of groups per study. To address this possibility, we extended our Study 2 simulations to obtain confidence intervals for $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1) for the five examples of group and shrinkage effects described in Table 2. For each example in Table 2, group sizes of 3, 5, 10, 15, 20 and 30 individuals, and samples of 15, 30, 45, and 60 groups were examined. The group sizes covered a wide range of group sizes in group-level studies, including studies with very small group sizes (e.g., Baer & Frese, 2003; Cole et al., 2013; Salanova, Agut, & Peiro, 2005) to those with a moderate to large number of persons per group (e.g., Berson et al., 2015; Schneider, White, & Paul, 1998; Zhang & Begley, 2011). Likewise, the values for the number of groups in a sample covered a considerable range of group-level studies including those with 30 or fewer groups (e.g., Dietz, Pugh, & Wiley, 2004; Luria & Yagil, 2008; Probst, 2015) to studies with more than 30 groups (e.g., Morrison et al., 2011; Naveh & Katz-Navon, 2015; Schneider et al., 2005).

We performed four repetitions of the examples in Table 2 with the 24 combinations of sample size and number of individuals per group. As shown across figures in the Supplemental Materials, bootstrapping results were very similar when resampling is based on both individuals and groups in comparison to only groups, with a slight advantage for only sampling groups. More specifically, the “true” population parameter was more frequently within the confidence interval limits for sampling only groups when the sample size is very small (i.e., 3 persons per group). These findings confirm what is known about the performance of bootstrapping methods with multilevel data (see Davison & Hinkley, 1997), and provide further support for our conclusions based on simulations with particular group sizes and numbers of groups per study.

General Discussion

The current research extends overall assessments of interrater agreement in two important ways. First, we presented a "central tendency approach" and procedures for assessing overall agreement, where we defined parameters denoting mean group agreement for the population of groups and presented a bootstrapping procedure for constructing confidence intervals around the mean population parameters for r_{WG}, AD, and ICC(1). Along with the mean statistic for the respective interrater agreement indices, we illustrated how confidence intervals for these indices provide additional information for assessing the extent of agreement across groups and how the range of values for a confidence interval can provide useful information for judging whether agreement is plausible in the population. In particular, we stressed how the range of values as determined by the confidence interval limits can be contrasted with proposed rules-of-thumb and ranges for the strength of an interrater agreement index to aid in data aggregation decisions and the interpretation of study findings. Thus, these tools provide researchers with more information concerning the quality of data from a set of groups, where data may be aggregated.

More specifically, when interpreting confidence intervals for mean interrater agreement, we advise researchers to make substantive interpretations of lower and upper limits relative to practical cutoff values for interrater agreement indices and ICC(1). As presented in our illustrations, this recommendation would call for focusing on the lower limits for r_{WG} and ICC(1) relative to rule-of-thumb cutoff values for these indices and the upper limit for AD relative to derived cutoff values for particular Likert-type scales (as presented in Smith-Crowe et al., 2013). When the lower limit for $r_{WG(J)}$ or ICC(1) is above the respective practical cutoff value, all plausible values for mean $r_{WG(J)}$ or ICC(1) in the population are acceptable; whereas, when the upper limit for mean AD is below a particular cutoff value (e.g., .85 on a 5-point scale), all plausible values for mean AD are acceptable. When a practical cutoff value is within the confidence interval, we recommend focusing on the percentage of values that are plausible and acceptable relative to those that are plausible but not practically acceptable. Finally, our illustrations also pointed to placing emphasis on the width of the confidence interval, the margin of error in estimating mean interrater agreement, as an index of precision.

Importantly, our argument to estimate interrater agreement as a population parameter and establish a confidence interval for the mean population parameter is also a call for constructively replicating group-level investigations to obtain a better understanding, over studies, of interrater agreement in the population. The confidence interval for mean interrater agreement from a primary group-level study is just one of many intervals that in the long run will capture mean interrater agreement with respect to a particular level of confidence. Across studies from different populations, knowledge of confidence intervals for mean interrater agreement will be useful in investigating where and why group-level study findings vary. In this regard, computing confidence intervals for mean interrater agreement in primary group-level studies would support meta-analyses and, according to Cumming and Finch (2005), “meta-analytic thinking focused on estimation” (p. 171).

Second, we extended assessments of overall interrater agreement by developing a new “matched difference approach” and procedures, using random group resampling. Here, matched differences between assessments of the respective r_{WG}, AD, and ICC(1) values for real and pseudo groups were computed, with the establishment of confidence intervals around these differences. Our findings indicate that this new approach and inference procedure will better inform researchers as to whether the difference between real and pseudo agreement across groups is sufficiently different from chance to conclude that the homogeneity within groups is real.

Notably, our results for the matched difference procedures indicated that ICC(1) is not superior to $r_{WG(J)}$ or $AD_{M(J)}$ for identifying differences between real and pseudo groups. Our results also pointed to situations where ICC(1) could produce a group effect in the absence of within-group agreement. Along with our findings, Woehr et al. (2015) noted that acceptable minimum values for ICC(1) are often based on personal observations or only a handful of articles in a particular domain, with some minimum values being quite small in magnitude (i.e., .01). Arguably, greater attention has been given to defining acceptable decision criteria for $r_{WG(J)}$ and $AD_{M(J)}$ based on theoretical, statistical, and psychometric considerations (e.g., see Smith-Crowe et al., 2014). These points coupled with the hybrid nature of ICC(1) as both an index of interrater reliability and interrater agreement raise questions about its usefulness in efforts to justify data aggregation.

It is important to note that the assessment of real versus pseudo agreement, via random group resampling approaches and procedures presented in this work and the work of other researchers, is empirical. As such, a finding supporting “pseudo agreement” does not necessarily take into account substantive reasons for a lack of variance across groups. Further, if one applies the consensus composition model (Chan, 1998), then arguably the threshold for aggregation is met by high levels of interrater agreement alone regardless of whether there are differences between groups. For instance, it could be the case that in a study of organizational climate for safety most or all organizations sampled have strong climates for safety (high $r_{WG(J)}$ values and low $AD_{M(J)}$ values) with similar, high means on a survey measure of climate. Such results could occur given that worker safety is highly regulated (e.g., in the U.S.), which may lead to a relative uniformity in organizations' focus on safety and workers' perceptions of organizational safety practices. Applying our matched differences approach in an instance like this one might indicate the presence of pseudo agreement. Yet, we would not presume that employees' agreement in their perceptions within a given organization is fake, or that they have pseudo safety climates.

Such a scenario highlights the need for researchers to apply RGR procedures thoughtfully with consideration of the circumstances of their data. Applying our matched difference procedures and interpreting RGR findings, like singular applications and interpretations of r_{WG}, AD, and ICC(1), should not be done without consideration of reasons for why and how group phenomena emerge in particular contexts. As such, researchers may need to consider how a variety of factors, including but not limited to, socio-political influences, group processes, and response biases affect the emergence of group phenomena and grouped data. In the case of using an interrater agreement procedure such as r_{WG} singularly or within an RGR analysis, the researcher needs to choose an appropriate null response distribution For instance, the situation where strong and uniform (across organizations) safety climates emerge as a result of government regulation and strict organizational safety practices, may necessitate the consideration of an "extreme groups" or "highly skewed" null response distribution (see Smith-Crowe et al., 2013) when applying and interpreting r_{WG} alone or within a matched difference RGR analysis.

The above discussion is not intended to imply that an assessment of the reliability of average ratings via ICC(2) (Bartko, 1976) or another means such as a test-retest reliability estimate for aggregated ratings (e.g., see Harter, Schmidt, Asplund, Killham, & Agrawal, 2010) is irrelevant. When assessing interrater agreement via any procedure, one may encounter situations with missing data. For instance, data may only be available for a percentage (e.g., 70%) of the members of each group. In such a situation, knowledge of the reliability of group means along with an assessment of interrater agreement may assist in informing a data aggregation decision. This point may be particularly important when there is a fair amount of missing data and where there are minimal differences between group means. Further, as discussed below, knowledge of the reliability of group means may be informative when correcting group-level correlations for measurement error (Bliese, 1998).

Additionally, our use of simulated data to evaluate the efficacy of our new inference procedures for assessing overall agreement called for the development of a new underlying model for multi-level data. This model and simulated data generation procedure proved efficacious in evaluating the accuracy of our proposed procedures for assessing overall interrater agreement. This model and procedures hold considerable promise for studying other interrater agreement problems as well as other phenomena in group-level studies. For instance, for primary group-level studies and meta-analytic studies with group-level data, Bliese (1998) and Burke, Landis, and Burke (2016), respectively, argued for the use of ICC(2) values in correcting group-level correlations. While these authors demonstrated the potential usefulness of such corrections with simulated and empirical data, their procedures did not allow for a more complete modeling of group level data and assessments of the accuracy of such reliability corrections. The availability of a model for use in generating multilevel data would assist in further evaluating the accuracy of such reliability corrections in primary group-level studies and meta-analytic studies with group-level data.

Another use of the underlying model for multi-level data and simulated data generation procedure would be extending the current work to further evaluate the efficacy of our procedures relative to other interrater agreement procedures not included in this investigation. For instance, based on the notion of random group resampling, Biemann, Ellwart, and Rack (2013) proposed an index referred to as r_{RG} that uses total variance in a sample of groups in the denominator of the r_{WG} item-level equation. Where a scale index is based on the average of the item indices. In effect, the use of total variance in the denominator of r_{RG} takes into account the variance of random groups and differences in item variances for a scale. Contrasting the performance of the r_{RG} procedure with our matched-difference procedures that involve random group resampling would be informative with respect to assessments of real versus pseudo agreement. In addition, extending the central tendency approach and procedures to include other indices such as the adjusted AD index (i.e., $AD_{M[J](adj)}$; Lohse-Bossenz, Kunina-Habenicht, & Kunter, 2013) would be useful in further evaluating the efficacy of central tendency procedures for assessments of overall interrater agreement.

In addition, research and practice on overall interrater agreement of groups might be meaningfully extended to include the consideration of a tolerance interval for interrater agreement (see Krishnamoorthy & Mathew, 2009; Young, 2010). The tolerance interval differs from confidence intervals in that confidence intervals bound the single-overall population parameter (i.e., mean interrater agreement) with some confidence, while a tolerance interval would bound the range of values that includes a specific proportion of the population. For instance, if we have a sample of 50 groups with a mean $r_{WG(J)}$ of .75 and a standard deviation of .15, we can establish a level of confidence (e.g., 90%) for a proportion of the population values (e.g., 90%, with respect to either a one-side or two-sided tolerance interval). The result for the lower one-sided interval for this example, under the assumption that the population is normally distributed, would be .52. This result would indicate that we can expect, with a high degree of confidence, that 90% of the values in the population are greater than .52. Consideration of tolerance intervals would call for greater understanding and more complete reporting of the nature of interrater agreement distributions in group-level studies, as our real datasets indicate that interrater agreement values such as $r_{WG(J)}$ and $AD_{M(J)}$ are likely non-normally distributed in the population. Open source software for the construction of tolerance intervals is widely available (e.g., see U.S. National Institute of Standards and Technology, 2013).

To facilitate and encourage the use of our approaches and procedures, we have developed a set of R functions, presented in Appendix B, for constructing bootstrapped confidence intervals around the mean population parameters for $r_{WG(J),}$ $AD_{M(J)}$, ICC(1), and their respective matched differences. Given the above discussion concerning the potential usefulness of ICC(2) in group-level studies, we also provide functions for constructing confidence intervals around the mean population parameter for ICC(2) and its matched difference. In addition, in Appendix B, we provide a function for simulating clustered data for one set of groups.

In sum, we provide “central tendency” and “matched difference” approaches and procedures for assessing overall agreement based on a sample of groups. Importantly, our approaches and procedures afford researchers the opportunity to move beyond assessments of interrater agreement within a sample of groups to assessments of interrater agreement for the population of groups. As such, our approaches and procedures advance understanding of the extent of agreement based on a sample of groups and, thus, can assist in better informing decisions to aggregate data and interpreting the generality of findings from group-level studies.

References

Abdelhadi, N., & Drach-Zahavy, A. (2011). Promoting patient care: Work engagement as a mediator between ward service climate and patient-centered care. *Journal of Advanced Nursing, 68*, 1276-1287.

American Psychological Association (2010). *Publication manual of the American Psychological Association.* Author: Washington, DC.

Aryee, S., Chen, Z. X., & Budhwar, P. S. (2004). Exchange fairness and employee performance: An examination of the relationship between organizational politics and procedural justice. *Organizational Behavior and Human Decision Processes*, *94*, 1-14.

Baer, M., & Frese, M. (2003). Innovation is not enough: Climates for initiative and psychological safety, process innovations, and firm performance. *Journal of Organizational Behavior*, 24: 45–68.

Bain, P., Mann, G., & Pirola-Merlo, A. (2001). The innovation imperative: The relationships between team climate, innovation, and performance in research and development teams. *Small Group Research, 32,* 55-73.

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. *Journal of Applied Psychology, 88*, 207-218.

Berson, Y., Da'as, R., & Waldman, D. (2015). How do leaders and their teams bring about organizational learning and outcomes? *Personnel Psychology, 68,* 79-108.

Biemann, T., Ellwart, T., & Rack, O. (2014). Quantifying similarity of team mental models: An introduction of the r_{WG} index. *Group Processes & Intergroup Relations, 17*, 125–140.

Binci, D. (2011). Climate for innovation and ICT implementation effectiveness: A missing link in Italian e-government projects. *International Journal of Public Administration, 34*, 49-53.

Bliese, P. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 4,* 355-373.

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. (1998). Group consensus and psychological well-being: A large field study. *Journal of Applied Social Psychology, 28,* 563-580.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein and S. W. J. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 349-381). San Francisco, CA: Jossey-Bass

Bliese, P. D., & Halverson, R. R. (2002). Using random group resampling in multilevel research: An example of the buffering effects of leadership climate. *The Leadership Quarterly, 13,* 53-68.

Boehm, S.A., Dwertmann, D.J.G., Bruch, H., Shamir, B. (2015). The missing link? Investigating organizational identity strength and transformational leadership climate as mechanisms that connect CEO charisma with firm performance. *The Leadership Quarterly, 26,* 156-171.

Boehm, S.A., Kunze, F., & Bruch, H. (2014). Spotlight on age-diversity climate: The impact of age-inclusive HR practices on firm-level outcomes. *Personnel Psychology, 67,* 667–704.

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

Brahm, T., & Kunze, F. (2012). The role of trust climate in virtual teams. *Journal of Managerial Psychology, 27,* 595 - 614

Bunderson, J. S. (2003). Team member functional background and involvement in management teams: Direct effects and the moderating role of power centralization. *Academy of Management Journal*, *46*, 458-474.

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.

Burke, M. I., Landis, R. S., & Burke, M. J. (2016). Estimating group-level relationships: General recommendations and considerations for the use of intraclass correlation coefficients. *Journal of Business and Psychology, 31*.

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.

Chen, X.P., Liu, D., & Portnoy, R. (2012). A multilevel investigation of motivational cultural intelligence, organizational diversity climate, and cultural sales: Evidence from U.S. real estate firms. *Journal of Applied Psychology, 97,* 93-106.

Chuang, C., & Liao, H. (2010). Strategic human resource management in service context: Taking care of business by taking care of employees and customers. *Personnel Psychology, 63,* 153–196.

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

Cole, M.S., Carter, M., & Zhang, Z. (2013). Leader-team congruence in power distance values and team effectiveness: The mediating role of procedural justice climate. *Journal of Applied Psychology, 98,* 962-973.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170-180.

Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap methods and their application.* Cambridge, UK: Cambridge University Press.

Dawson, J. F., González-Romá, 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.

Dietz, J., Pugh, S. D., & Wiley, J. (2004). Service climate effects on customer attitudes: An examination of boundary conditions. *Academy of Management Journal*, 47: 81-92.

de Jong, K., de Ruyter, K., & Lemmink, J. (2005). Service climate in self-managing teams: Mapping the linkage of team member perceptions and service performance outcomes in a business-to-business setting. *Journal of Management Studies, 42*, 1593-1620.

Dong, Y., Liao, H., Chuang, A., Zhou, J., & Campbell, E. M. (2015). Fostering employee service creativity: Joint effects of customer empowering behaviors and supervisory empowering leadership. *Journal of Applied Psychology, 100,* 1364-1380.

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.

Earley, C. P., & Mosakowski, E. (2000). Creating hybrid team cultures: An empirical test of transnational team functioning. *Academy of Management Journal, 43(1),* 26-49.

Eisenbeiss, S., Knippenberg, D.V. Boerner, S. (2008). Transformational leadership and team innovation: Integrating team climate principles. *Journal of Applied Psychology,* 93, 1438-1446.

Erez, M., Lisak, A., Harush, R., Glikson, E., Nouri, R., & Shokef, E. (2013). Going global: Developing management students' cultural intelligence and global identity in culturally diverse virtual teams. *Academy of Management Learning & Education, 12*, 330-355.

Fahr, J.L, Lee, C., & Fahr, C.I. (2010). Task conflict and team creativity: A question of how much and when. *Journal of Applied Psychology, 95,* 1173-1180.

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

Gil, F., Rico, R., Alcover, C.M., & Barrasa, M. (2005). Change-oriented leadership, satisfaction and performance in work groups: Effects of team climate and group potency, *Journal of Managerial Psychology, 20,* 312-328.

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

Gonzalez-Roma, V., & Hernandez, A. (2014). Climate uniformity: Its influence on team communication quality, task conflict, and team performance. *Journal of Applied Psychology, 99,* 1042-1058.

Griffith, J. (2006). A compositional analysis of the organizational climate-performance relation: Public schools as organizations. *Journal of Applied Social Psychology, 36*, 1848-1880.

Gupta, M., Uz, I., Esmaeilzadeh, P., Noboa, F., Mahrous, A.A., Kim, E., Miranda, G., Tennant, V.M., Chung, S., Azam, A., Peters, A., Iraj, H., Bautista, V.B, Kulikova, I. (2018). Do cultural norms affect social network behavior inappropriateness? A global study. *Journal of Business Research*, *85*, 10-22.

Harter, J.K., Schmidt, F.L., Asplund, J. W., Killham, E.A., & Agrawal, S. (2010). Causal impact of employee work perceptions on the bottom line of organizations. *Perspectives on Psychological Science, 5*, 378-389.

Hofmann, D. A., & Mark, B. (2006). An investigation of the relationship between safety climate and medication errors as well as other nurse and patient outcomes. *Personnel Psychology*, 59: 847-869.

Hui, C. H., Chiu, W. C. K., Yu, P. L. H., Cheng, K., & Tse, H. M. (2007). The effects of service climate and the effective leadership behavior of supervisors on frontline employee service quality: A multi-level analysis. *Journal of Occupational and Organizational Psychology, 80,* 151–172.

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

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

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

Jiang, K, Chuang, C., & Chiao, Y. (2015). Developing collective customer knowledge and service climate: The interaction between service-oriented high-performance work systems and service leadership. *Journal of Applied Psychology, 100*, 1089–1106.

Katz-Navon, T., Naveh, E., & Stern, Z. (2005). Safety climate in healthcare organizations: A multidimensional approach. *Academy of Management Journal, 48*, 1075–1089.

Kirkman, B. L., Tesluk, P. E., & Rosen, B. (2001). Assessing the incremental validity of team consensus ratings over aggregation of individual-level data in predicting team effectiveness. *Personnel Psychology, 54,* 645-667.

Koene, B. A. S., Vogelaar, A. L. W., & Soeters, J. L. (2002). Leadership effects on organizational climate and financial performance: Local leadership effect in chain organizations. *The Leadership Quarterly, 13*, 193–215.

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: Foundations, Extensions, and New Directions* (pp. 3 – 90). San Francisco, CA: Jossey-Bass.

Krishnamoorthy, K., & Mathew, T. (2009). *Statistical tolerance regions: Theory, applications, and computation*. Hoboken, NJ: John Wiley & Sons.

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

Liu, D., Hernandez, M., & Wang, L. (2014). The role of leadership and trust in creating structural patterns of team procedural justice: A social network investigation. *Personnel Psychology, 67,* 801-846.

Lohse-Bossenz, H., Kunina-Habenicht, O., & Kunter, M. (2013). Estimating within-group agreement in small groups: A proposed adjustment for the average deviation index. *European Journal of Work and Organizational Psychology*.

Ludtke, 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., & Yagil, D. (2008). Procedural justice, ethical climate and service outcomes in restaurants. *International Journal of Hospitality Management*, 27: 276-283.

MacCormick, J. S., & Parker, S. K. (2010). A multiple climates approach to understanding business unit effectiveness. *Human Relations, 63*, 1171–1806.

Mannix, E. A. (2001). The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. *Academy of Management Journal, 44(4),* 238-251.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

McKay, P. F., Avery, D. R., & Morris, M. A. (2009). A tale of two climates: Diversity climate from subordinates' and managers' perspectives and their role in store unit sales performance. *Personnel Psychology, 62,* 767–791.

Menges, J.I., Walter, F., Vogel, B., & Bruch, H. (2011). Transformational leadership climate:

Performance linkages, mechanisms, and boundary conditions at the organizational level. *The Leadership Quarterly, 22,* 893-909.

Morrison, E.W., Wheeler-Smith, S.L., & Kamdar, D. (2011). Speaking up in groups: A cross-level study of group voice climate and voice. *Journal of Applied Psychology, 96*, 183–191.

Myer, A. T., Thoroughgood, C. N., & Mohammed, S. (2016). Complementary or competing climates? Examining the interactive effect of service and ethical climates on company-level financial performance. *Journal of Applied Psychology, 101,* 1178-1190.

National Institute of Standards and Technology (NIST). (2013). *NIST/SEMATECH e-Handbook of Statistical Methods*. Washington, D.C: NIST.

Naveh, E., & Katz-Navon, T. (2015). A longitudinal study of an intervention to improve road safety climate: Climate as an organizational boundary spanner. *Journal of Applied Psychology, 100,* 216-226.

Neal, A., & Griffin, M. A. (2006). A longitudinal study of the relationships among safety climate, safety behavior,and accidents at the individual and group levels. *Journal of Applied Psychology*, 91: 946-953.

Neal, A., West, M.A., & Patterson, M.G. (2005). Do organizational climate and competitive strategy moderate the relationship between human resource management and productivity? *Journal of Management, 31,* 492-512.

Pasisz, D. J., & Hurtz, G. M. (2009). Testing for between-group differences in within-group interrater agreement. *Organizational Research Methods, 12,* 590-613.

Patterson, M.G, West, M.A., Shackleton, V.J., Dawson, J.F., Lawthom, R., Maitlis, S., Robinson, D.L., Wallace, A.S. (2002). Validating the organizational climate measure: links to managerial practices, productivity and innovation. Journal of Organizational. Behavior, 26, 379–408.

Probst, T. M. (2015). Organizational safety climate and supervisor safety enforcement: multilevel explorations of the causes of accident underreporting. *Journal of Applied Psychology, 100,* 1899–1907.

Ren, S., Lai, H., Tong, W, Aminzadeh, M., Hou, X, & Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, 37, 1487-1498.

Riordan, C.M., Vandenberg, R.J., & Richardson, H.A. (2005). Employee involvement climate and organizational effectiveness. *Human Resource Management, 44,* 471-488.

Salanova, M., Agut, S., & Peiro, J. M. (2005). Linking organizational resources and work engagement to employee performance and customer loyalty: The mediation of service climate. *Journal of Applied Psychology*, 90: 1217-1227.

Schneider, B., Ehrhart, M. G., Mayer, D. M., Saltz, J. L., & Niles-Jolly, K. (2005). Understanding organization-customer links in service settings. *Academy of Management Journal*, 48: 1017-1032.

Schneider, B., White, S. S., & Paul, M. C. (1998). Linking service climate and customer perceptions of service quality: Test of a causal model. *Journal of Applied Psychology*, 83: 150-163.

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

Simons, T., & Roberson, Q. (2003). Why managers should care about fairness: The effects of aggregate justice perceptions on organizational outcomes. *Journal of Applied Psychology*, 88: 432-443.

Smith, K. G., Collins, C. J., & Clark, K. D. (2005). Existing knowledge, knowledge creation capability, and the rate of new product introduction in high-technology firms. *Academy of Management Journal, 48*, 346-357.

Smith-Crowe, K., Burke, M. J., Cohen, A., & Doveh, E. (2014). Statistical significance criteria for the r_{WG} and average deviation interrater agreement indices. *Journal of Applied Psychology, 99,* 239–261

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.

Sowinski, D. R., Fortmann, K. A., & Lezotte, D. V. (2008). Climate for service and the moderating effects of climate strength on customer satisfaction, voluntary turnover, and profitability. *European Journal of Work and Organizational Psychology, 17,* 73-88.

Susskind, A. M., Kacmar, K. M., & Borchgrevink, C. P. (2003). Customer service providers' attitudes relating to customer service and customer satisfaction in the customer-server exchange. *Journal of Applied Psychology, 88*, 179-187.

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.

Towler, A., Lezotte, D. V., & Burke, M. J. (2011). An examination of the service climate-firm performance chain: The role of customer retention. *Human Resource Management, 50*, 391-406.

Wallace, J. C., & Chen, G. (2006). A multilevel integration of personality, climate, self-regulation, and performance. *Personnel Psychology, 59*, 529–557.

Wallace, J.C., Johnson, P.D., Mathe, K., & Paul, J (2011). Structural and psychological empowerment climates, performance, and the moderating role of shared felt accountability: A managerial perspective. *Journal of Applied Psychology, 96,* 840-850.

Woehr, D. J., Loignon, A. C., & Schmidt, P. J. (2015a). Aggregation aggravation: The fallacy of the wrong level revisited. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 311-326). New York, NY: Routledge.

Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015b). Justifying aggregation with consensus-based constructs: A review and Examination of cutoff values for common aggregation indices. *Organizational Research Methods.*

Young, D. S. (2010). Tolerance: An R package for estimating tolerance intervals. *Journal of Statistical Software, 36*, 1–39.

Table 1

Study 1 Results for the SIM1A, SIM1B, and TMCTP Datasets

		$r_{WG(J)}$	$AD_{M(J)}$	ICC(1)
SIM1A				
1000 Monte Carlo Samples	2.5 Percentile	0.392	0.835	0.142
	50 Percentile	0.528	0.910	0.313
	97.5 Percentile	0.647	0.994	0.477
Single Selected Sample	Parameter Estimate[a]	0.596	0.868	0.402
	95% Confidence Interval	0.467, 0.713	0.780, 0.947	0.189, 0.563
SIM1B				
1000 Monte Carlo Samples	2.5 Percentile	0.695	0.320	0.736
	50 Percentile	0.816	0.462	0.846
	97.5 Percentile	0.911	0.613	0.922
Single Selected Sample	Parameter Estimate[a]	0.877	0.421	0.875
	95% Confidence Interval	0.800, 0.937	0.290, 0.551	0.781, 0.932
TMCTP				
Task Conflict	Parameter Estimate	0.809	0.824	0.186
	95% Confidence Interval	0.761, 0.852	0.755, 0.899	0.043,0.332
Team Identity	Parameter Estimate	0.949	0.485	0.244
	95% Confidence Interval	0.936, 0.959	0.441, 0.535	0.118,0.357

Note. We consider the median values of the 1000 Monte Carlo samples to be the "true" population values for SIM1A and SIM1B, respectively. [a]The parameter estimate is the estimated value from the single selected bootstrap sample. The Task Conflict and Team Identity measures were from time point 3 out of four time points.

Table 2

Parameters Specific to Each Simulated Dataset in Study 2

Dataset	Group Effect		Shrinkage	
SIM2A	No Group Effect	σ_α=0.0	Shrinkage	θ_1=0.1, θ_2=0.8
SIM2B	Larger Group Effect	σ_α=2.5	No Shrinkage	θ_1=1.0, θ_2=0.0
SIM2C	Larger Group Effect	σ_α=2.5	Shrinkage	θ_1=0.1, θ_2=0.8
SIM2D	Small Group Effect	σ_α=1.0	Shrinkage	θ_1=0.1, θ_2=0.8
SIM2E	Large Group Effect	σ_α=2.0	Anti-Shrinkage	θ_1=2.0, θ_2=0.0

Note. Group effect refers to the extent that there is variance between groups. Shrinkage refers to the contraction of the distribution of responses, such that the range is narrower. Anti-shrinkage refers to an expansion of the distribution of responses, such that the range is wider. Following from the model presented in the Appendix, θ_2=0.0 implies v_k=θ_1=CONSTANT=2. We purposely selected a constant greater than 1.

Table 3

Study 2 Results for the TMCTP Dataset

			Percentile	
		2.5	50	97.5
	Task Conflict Measure			
$r_{WG(J)}$	Bootstrap Sample	0.769	0.827	0.873
(0.824)	Matched, Shuffled Sample	0.769	0.820	0.864
	Difference	**-0.027**	**0.006**	**0.037**
$AD_{M(J)}$	Bootstrap Sample	0.710	0.788	0.877
(0.792)	Matched, Shuffled Sample	0.716	0.795	0.883
	Difference	**-0.062**	**-0.007**	**0.052**
ICC(1)	Bootstrap Sample	-0.118	-0.006	0.123
(0.000)	Matched, Shuffled Sample	-0.098	-0.002	0.119
	Difference	**-0.172**	**-0.003**	**0.160**
	Team Identity Measure			
$r_{WG(J)}$	Bootstrap Sample	0.936	0.950	0.959
(0.949)	Matched, Shuffled Sample	0.920	0.938	0.951
	Difference	**0.003**	**0.012**	**0.022**
$AD_{M(J)}$	Bootstrap Sample	0.441	0.487	0.535
(0.485)	Matched, Shuffled Sample	0.474	0.529	0.590
	Difference	**-0.084**	**-0.043**	**-0.006**
ICC(1)	Bootstrap Sample	0.118	0.239	0.357
(0.244)	Matched, Shuffled Sample	-0.100	-0.002	0.122
	Difference	**0.069**	**0.238**	**0.396**

Note. The numbers in parentheses in the first column are the observed $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1) values for the sample prior to bootstrapping. The $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1) differences are the differences between a given bootstrap sample and its matched, shuffled sample. The 2.5 percentile of the difference values is the lower bound of the 95% confidence interval for the respective statistic and the 97.5 percentile is the upper bound. The 50 percentile is the median of the *B* values. The Task Conflict was from time 3 and the Team Identity from time 2.

Table 4

Study 2 Results for the SIM2 Datasets

		Percentile		
		2.5	50	97.5
	SIM2A			
$r_{WG(J)}$	Bootstrap Sample	0.691	0.769	0.835
(0.769)	Matched, Shuffled Sample	0.724	0.804	0.862
	Difference	**-0.075**	**-0.032**	**0.013**
$AD_{M(J)}$	Bootstrap Sample	0.683	0.756	0.827
(0.755)	Matched, Shuffled Sample	0.708	0.780	0.851
	Difference	**-0.055**	**-0.023**	**0.002**
ICC(1)	Bootstrap Sample	-0.013	0.049	0.116
(0.051)	Matched, Shuffled Sample	-0.039	-0.001	0.046
	Difference	**-0.024**	**0.048**	**0.131**
	SIM2B			
$r_{WG(J)}$	Bootstrap Sample	0.653	0.756	0.859
(0.754)	Matched, Shuffled Sample	0.000	0.025	0.077
	Difference	**0.611**	**0.728**	**0.840**
$AD_{M(J)}$	Bootstrap Sample	0.343	0.480	0.616
(0.480)	Matched, Shuffled Sample	1.456	1.553	1.649
	Difference	**-1.285**	**-1.072**	**-0.866**
ICC(1)	Bootstrap Sample	0.696	0.794	0.876
(0.795)	Matched, Shuffled Sample	-0.039	0.001	0.046
	Difference	**0.686**	**0.791**	**0.885**
	SIM2C			
$r_{WG(J)}$	Bootstrap Sample	0.906	0.946	0.976
(0.946)	Matched, Shuffled Sample	0.000	0.027	0.098
	Difference	**0.833**	**0.917**	**0.965**

$AD_{M(J)}$ (0.253)	Bootstrap Sample	0.164	0.254	0.346
	Matched, Shuffled Sample	1.404	1.549	1.660
	Difference	**-1.472**	**-1.292**	**-1.084**
ICC(1) (0.934)	Bootstrap Sample	0.888	0.932	0.967
	Matched, Shuffled Sample	-0.037	0.000	0.045
	Difference	**0.870**	**0.932**	**0.986**
	SIM2D			
$r_{WG(J)}$ (0.897)	Bootstrap Sample	0.866	0.897	0.927
	Matched, Shuffled Sample	0.091	0.208	0.359
	Difference	**0.525**	**0.690**	**0.821**
$AD_{M(J)}$ (0.561)	Bootstrap Sample	0.484	0.563	0.637
	Matched, Shuffled Sample	1.091	1.196	1.303
	Difference	**-0.808**	**-0.630**	**-0.478**
ICC(1) (0.771)	Bootstrap Sample	0.674	0.765	0.840
	Matched, Shuffled Sample	-0.037	-0.001	0.046
	Difference	**0.667**	**0.765**	**0.852**
	SIM2E			
$r_{WG(J)}$ (0.363)	Bootstrap Sample	0.249	0.364	0.486
	Matched, Shuffled Sample	0.000	0.024	0.074
	Difference	**0.219**	**0.339**	**0.465**
$AD_{M(J)}$ (1.079)	Bootstrap Sample	0.918	1.080	1.226
	Matched, Shuffled Sample	1.466	1.549	1.619
	Difference	**-0.660**	**-0.469**	**-0.297**
ICC(1) (0.383)	Bootstrap Sample	0.261	0.378	0.496
	Matched, Shuffled Sample	-0.038	0.000	0.043
	Difference	**0.255**	**0.380**	**0.505**

Note. The numbers in parentheses in the first column are the observed $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1) values for the single sample drawn from each *B*=1000 simulations. The $r_{WG(J)}$, $AD_{M(J)}$, and ICC(1) differences are the differences between a given bootstrap sample and its matched, shuffled sample. The 2.5 percentile of the difference values is the lower bound of the 95% confidence interval for the respective statistic and the 97.5 percentile is the upper bound. The 50 percentile is the median of the *B* values.

Appendix: A New Model for Generating Multilevel Data

Here we explain the model we developed and used to generate multilevel data. To begin, Y_{ijk} denotes the response of the *i*'th individual for item *j* and is equal to 1,…*A*, where *A* denotes the number of response options on a Likert scale. Thus, Y_{ijk} is what the model is designed to produce. The number of responses within a single group *k* is denoted by n_k, *K* is the total number of groups and *J* is the total number of items. We assume that the observed value is generated by an underlying continuous response defined as follows:

$$X_{ijk},\quad \begin{matrix} i = 1, 2, \dots, n_k \\ j = 1, 2, \dots, J \\ k = 1, 2, \dots, K \end{matrix} \qquad (A1)$$

We further assume that the underlying response is a result of two additive variables: a latent variable *F*, which characterizes individual *i* nested in group *k*, and a deviation for each item *j*. Thus,

$$X_{ijk} = \beta_j F_{ik} + \varepsilon_{ijk}. \qquad (A2)$$

Similar to the factor analysis model, F_{ik} is assumed to be normally distributed and β_j is the loading on *F*, which we set to be equal for all *J* items. The error/uniqueness term is

$$\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2) iid, \qquad (A3)$$

where σ_ε is the standard deviation of the uniqueness.

There are two possibilities for responses of subjects in the same group. That is, they are independent or they share some common characteristics. In the latter case we model F_{ik} as

$$F_{ik} = v_k \delta_{ik} + \alpha_k \qquad (A4)$$

$$\delta_{ik} \sim N\left(0, \sigma_{\delta}^{2}\right) iid, \qquad (A5)$$

Here, delta expresses the individual's random effect within group k and α_k represents the random shift effect of clustering on F_{ij}. That is, common to all members in group k their F_{ij}, is shifted by α_k, relative to the general mean over all groups. Where α_k is distributed as

$$\alpha_k \sim N(0, \sigma_{\alpha}^{2}) iid. \qquad \text{(A6)}$$

The effect of this parameter will be reflected mainly in observed ICC(1) values because varying α increases the variance between groups, but does not change the within group variance. The parameter α resembles the random intercept in HLM models, which expresses the differences in the intercepts between the higher level categories, but does not account for differences in the slopes of the explanatory variables.

A model that includes only the parameter α to account for group differences is not rich enough for capturing multi-group data structures with different combinations of agreement indices and ICC(1) values. Therefore, the random component v_k was included and was modelled as distributed Beta with the parameters θ_1, θ_2, ω_1, ω_2. Its inclusion is analogous to the extension to random regression coefficients, in the random coefficient models.

$$v_k \sim \theta_1 + \theta_2 * BETA(\omega_1, \omega_2) iid. \qquad (A7)$$

The parameters θ_1 and θ_2 are location and scale parameters, respectively. The θ_1 shifts and the θ_2 shrinks, or expands, the random effect generated by the beta distribution. We used the Beta distribution since its values are in the range of 0 to 1.

In summary, the model includes several components that form the underlying response of individual *i* in the *k*'th group: the random part of the group α_k, and the random variable which is the product of δ_{ik} and v_k. Each of the two group level components (α_k and v_k) affect the within group agreement, determining the r_{WG} value of the group. When θ_1=1 and θ_2=0, there is no shrinkage, v_k =1, and F_{ik}=δ_{ik}. Then, according to the model, if observations are independent, then

$$X_{ijk} = \beta\delta_{ik} + \varepsilon_{ijk}. \qquad (A8)$$

If observations are dependent, then

$$X_{ijk} = \beta(v_k\delta_{ik} + \alpha_k) + \varepsilon_{ijk}. \qquad (A9)$$

The respective variances for independent and dependent observations are then

$$Var\big(X_{ijk}\big) = \beta^2\sigma_\delta^2 + \sigma_\varepsilon^2 \qquad (A10)$$

$$Var\big(X_{ijk}\big) = \beta^2(V(v_k\delta_{ik}) + \sigma_\alpha^2) + \sigma_\varepsilon^2, \qquad (A11)$$

where

$$Var(v_k\delta_{ik}) = (\theta_2)^2 * \left(\frac{\omega_1\omega_2}{(\omega_1+\omega_2)^2 * (\omega_1+\omega_2+1)}\right) * \sigma_\delta^2 + \left(\theta_1 + \theta_2 * \left(\frac{\omega_1}{\omega_1+\omega_2}\right)\right)^2 * \sigma_\delta^2. \quad (A12)$$

Equation A13 was derived as follows. The variance of the product of two independent variables U, V is by definition

$$Var(UV) = E\{(UV)^2\} - \{E(UV)\}^2. \qquad (A13)$$

This equation can also be stated as

$$Var(U)Var(V) + Var(U)E(V)^2 + Var(V)E(U)^2. \qquad (A14)$$

In our case U=δ_{ik} with expected value equal to 0, variance=σ_δ^2,

$$V = v_k = \theta_1{+}\theta_2 * BETA(\omega_1, \omega_2). \qquad (A15)$$

The expected value of a Beta variable is

$$\frac{\omega_1}{\omega_1+\omega_2} \qquad (A16)$$

and its variance is

$$\frac{\omega_1\omega_2}{(\omega_1+\omega_2)^2 * (\omega_1+\omega_2+1)}. \qquad (A17)$$

Therefore,

$$E(v_k) = \theta_1 + \theta_2 * \left(\frac{\omega_1}{\omega_1 + \omega_2}\right) \qquad (A18)$$

and

$$Var(v_k) = (\theta_2)^2 * \left(\frac{\omega_1\omega_2}{\omega_1 + \omega_2 + 1}\right). \qquad (A19)$$

One should distinguish between the variance of the response when we consider the whole population of groups versus the variance within a group, because, unlike the latter, considering the whole population of groups entails the variance within each group as well as the variance between groups. We denote the variance within groups as $\sigma^2(X_{ijk} | \alpha_k, v_k)$, referring to the variance within group *k* for which the "group shift effect" was α_k and the "group shrinkage effect" was v_k. In other words, α_k represents the extent to which the entire distribution of responses shifts on the Likert scale, and v_k represents the extent to which the range of the distribution of responses increases or decreases within the group. The distribution of responses conditional on α_k and v_k is $N(\beta\alpha_k, \beta^2 v_k^2 \sigma_\delta^2 + \sigma_\varepsilon^2)$. α_k and v_k are fixed with group *k*, σ_δ^2 is the variance between individuals in the group, and σ_ε^2 is the error variance.

The observed values are the discrete Likert scale Y's. We denote the expected probabilities of the ordinal responses conditional on α_k and v_k as $P_1(\alpha_k, v_k),\ldots,P_A(\alpha_k, v_k)$.

$$P_1(\alpha_k, v_k) = P\left(Y_{ijk} \le \Phi^{-1}(PM_1) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}\right) \qquad (A20)$$

Equation A20 can be rewritten as follows:

$$P_1(\alpha_k, v_k) = P\left(\frac{Y_{ijk} - \beta\alpha_k}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}} \le \frac{\Phi^{-1}(PM_1) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}} - \frac{\beta\alpha_k}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}}\right) \qquad (A21)$$

$$P_1(\alpha_k, v_k) = \Phi\left(\frac{\Phi^{-1}(PM_1) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}} - \frac{\beta\alpha_k}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}}\right) \qquad (A22)$$

$\Phi^{-1}(p)$ denotes the *p'th* quantile of the standard normal distribution.

Note that if the data are independent and there is no shrinkage, then α_k is equal to 0 and v_k is equal to 1. Hence,

$$P_1(\alpha_k, v_k) = \Phi\left(\frac{\Phi^{-1}(PM_1) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}}{\sqrt{\beta^2 * 1 * \sigma_\delta^2 + \sigma_\varepsilon^2}} - \frac{\beta * 0}{\sqrt{\beta^2 * 1 * \sigma_\delta^2 + \sigma_\varepsilon^2}}\right), \qquad (A23)$$

which can be rewritten as

$$P_1(\alpha_k, v_k) = \Phi\big(\Phi^{-1}(PM_1)\big) = PM_1. \qquad (A24)$$

The general term for the probabilities of Y with 1>a>A are

$$P_a(\alpha_k, v_k) = \Phi\left(\frac{\Phi^{-1}(PM_a) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}} - \frac{\beta\alpha_k}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}}\right) - \Phi\left(\frac{\Phi^{-1}(PM_a) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}} - \frac{\beta\alpha_k}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}}\right) \quad (A25)$$

and for a=A,

$$P_A(\alpha_k, v_k) = 1 - \Phi\left(\frac{\Phi^{-1}(PM_A) * \sqrt{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}} - \frac{\beta\alpha_k}{\sqrt{\beta^2 v_k^2\sigma_\delta^2 + \sigma_\varepsilon^2}}\right). \qquad (A26)$$

The mean of the ordinal item variances within group with the random effects α_k and v_k is therefore expected to be

$$Var\big(Y_{ijk}|\alpha_k, v_k\big) = \textstyle\sum_{a=1}^{A} a^2 P_a(\alpha_k, v_k) - (\sum_{a=1}^{A} aP_a(\alpha_k, v_k))^2. \qquad (A27)$$

Appendix B: R Functions for Simulating Data for a Set of Groups and Constructing Bootstrapped Confidence Intervals

```
########################################################################
#  This appendix includes:
# - SimDat_cr: A function for simulating one set of groups (clustered) data
# - Creating 5 examples of clustered data (with the parameters presented
#   in Table 2 of the article)
# - Functions for constructing bootstrapped confidence
#   intervals around the mean population parameters for
#   rwg, AD, ICC(1), ICC(2), and their respective “matched difference”
########################################################################

########################################################################
# Multilevel and reshape R libraries should be installed
#    (if not already installed in the user's R libraries) prior
#    to using our code
# These are R libraries and hence are installed from within R
#    the usual way.
# One of the options to install these libraries is as follows:
#    The computer needs to be connected to the internet,
#    and R running. Then type within R:
#
#      install.packages("multilevel")
#      install.packages("reshape")
#
# reshape library is needed for simulating data (part 1 of the code)
# multilevel library is needed for calculating RWG AD and ICC
#   (part 2 of the code)
################################################################

#SimDat_cr function:
#   Simulate one set of groups data
#   and return a data set of (gsize X gnum) rows
#   and nitems+2 columns of data
#   - 1st column displays id
#   - 2nd column displays the cluster(=group) number
#   - Last nitems columns include the data of the items
################################################################

#Note: Fix the seed before you start the simulations
#      and load libraries (if not already loaded)
#  library(reshape)

#Arguments:
#nitems  - Number of items to simulate
#pvec    - Vector of probabilities for the ordinal variable.
```

```
#FLoad   - factor loadings (this function simulates data
#         with equal loadings for all nitems
#gsize   - Simulated group size(number of subjects per cluster/group)
#gnum    - number of groups/clusters
#sd_eps  - the standard deviation of the uniqueness
#sd_del  - is the standard deviation of  the delta part of the
#         latent construct F, when the data are not clustered
#sd_alph - is the standard deviation of  the alpha part of the
#         latent construct F.
#         This is the shift effect of clustering on F_ij
#omega1  - 1st shape parameter of the beta distribution
#omega2  - 2nd shape parameter of the beta distribution
#Theta1  - shift of the beta distribution parameter
#Theta2  - shrink of the beta distribution parameter

#Value
#   returns a data set of (gsize X gnum) rows of data
#   and nitems+1 columns
#   1st nitems columns includes the items data and the last column
#   is the cluster number

SimDat_cr<-function(nitems, pvec, FLoad, gsize, gnum,
              sd_eps, sd_del, sd_alph,
              omega1, omega2, Theta1, Theta2
)
{
  marginal<-cumsum(pvec[1:(length(pvec)-1)])
              # CDF of the ordinal variable probability
              # marginal cumulative probability defining the
              # marginal distribution of the ordinal variable

  a<-length(marginal)+1    #number of response options
  beta<-rep(FLoad,nitems)  # function simulates data
                      # with equal loadings for all nitems

  SD_x_I <- sqrt((FLoad**2)*(sd_del**2)+sd_eps**2) #std of independent X
  cut_p <-qnorm(marginal)*SD_x_I  #cut points of independent X

  #Generate a sample of clustered data
  eps <- rnorm(nitems*gsize*gnum,0,sd_eps)
  delta <- rnorm(gsize*gnum,0,sd_del)
  alpha <- rnorm(gnum,0,sd_alph)
  gamma <- Theta1+Theta2*rbeta(gnum,omega1,omega2)

  #Create the Xij data
  dat <- expand.grid(
         item=factor(1:nitems),
         id=factor(1:gsize),
         cluster=factor(1:gnum))

  SimDat<- data.frame(dat,betaM=model.matrix(~ 0+ item, dat)%*%beta,
              epsM=eps)
  #loop
  SimDat$alphaM<-rep(NA,nitems*gsize*gnum)
  SimDat$gammaM<-rep(NA,nitems*gsize*gnum)
  i_count<-0

  for(i in 1:gnum){
   #cat("i ",i,"\n")
   SimDat$alphaM[SimDat$cluster==i]<-alpha[i]
   SimDat$gammaM[SimDat$cluster==i]<-gamma[i]
   for(j in 1:gsize){
    i_count<-i_count+1
    SimDat$deltaM[(SimDat$cluster==i & SimDat$id==j)]<-delta[i_count]
   }
  }
  SimDat$F_C <- SimDat$gammaM*SimDat$deltaM +
              SimDat$alphaM #clustered observations F (factor)
  SimDat$xM_C <- SimDat$betaM*SimDat$F_C+SimDat$epsM  #clustered x

  #Create Y_ijk (ordinal variables)
  SimDat$yM_C<-as.numeric(cut(SimDat$xM_C,c(min(SimDat$xM_C)-1,
             cut_p,max(SimDat$xM_C)+1))) #ordinalize data

  #Create data by items
  items_C<-SimDat[SimDat[,"item"]==1,c("id","cluster") ]

  for(i in 1:nitems){
   x<-SimDat[SimDat[,"item"]==i,c("yM_C") ]
   nm_x<-c(names(items_C),paste("item",i,sep=""))
   items_C<-cbind(items_C,x)
   names(items_C)<-nm_x
  }
  #rename cluster as group)
  dat <- rename(items_C, c(cluster="group"))
  row.names(dat)<-NULL
  dat
}
```

```
#########################################################
# Simulating the 5 examples (Table 2 of the manuscript)
#########################################################

#Load libraries (if not already loaded)
library(reshape)

#Example 1: sd_alph=0 => no random effect + shrinkage
set.seed(5666223)
Ex1_dat1<-SimDat_cr(nitems =6, pvec=c(.2,.2,.2,.2,.2),
    FLoad=0.7, gsize = 10, gnum = 50,
    sd_eps = 0.2, sd_del = 1, sd_alph = 0,
    omega1 = 1, omega2=1,
    Theta1 = 0.1, Theta2=.8)

#Example 2: Theta2=0 => no shrinkage + random effect
set.seed(5666223)
Ex2_dat1<-SimDat_cr(nitems =6, pvec=c(.2,.2,.2,.2,.2),
    FLoad=0.7, gsize = 10, gnum = 50,
    sd_eps = 0.2, sd_del = 1, sd_alph = 2.5,
    omega1 = 1, omega2=1,
    Theta1 = 1, Theta2=0)

#Example 3: sd_alph=2.5 Theta1 = 0.1 Theta2=.8 => random effect+shrinkage
set.seed(5666223)
Ex3_dat1<-SimDat_cr(nitems =6, pvec=c(.2,.2,.2,.2,.2),
    FLoad=0.7, gsize = 10, gnum = 50,
    sd_eps = 0.2, sd_del = 1, sd_alph = 2.5,
    omega1 = 1, omega2=1,
    Theta1 = 0.1, Theta2=.8)

#Example 4: sd_alph=1 Theta1 = 0.1 Theta2=.8 => smaller random effect+shrinkage
set.seed(5666223)
Ex4_dat1<-SimDat_cr(nitems =6, pvec=c(.2,.2,.2,.2,.2),
    FLoad=0.7, gsize = 10, gnum = 50,
    sd_eps = 0.2, sd_del = 1, sd_alph=1,
    omega1 = 1, omega2=1,
    Theta1 = 0.1, Theta2=.8)

#Example 5: random effect+ anti - shrinkage(shrinkage=2)
set.seed(5666223)
Ex5_dat1<-SimDat_cr(nitems =6, pvec=c(.2,.2,.2,.2,.2),
    FLoad=0.7, gsize = 10, gnum = 50,
    sd_eps = 0.2, sd_del = 1, sd_alph=2,
    omega1 = 1, omega2=1,
    Theta1 = 2, Theta2=0)
```

```
#write examples into csv files
write.csv(Ex1_dat1,"Ex1_dat1.csv",row.names=FALSE)
write.csv(Ex2_dat1,"Ex2_dat1.csv",row.names=FALSE)
write.csv(Ex3_dat1,"Ex3_dat1.csv",row.names=FALSE)
write.csv(Ex4_dat1,"Ex4_dat1.csv",row.names=FALSE)
write.csv(Ex5_dat1,"Ex5_dat1.csv",row.names=FALSE)

################################################################################
################################################################################

#functions for constructing bootstrapped confidence
#   intervals around the mean population parameters of
#   rwg AD ICC and their “matched difference”
#If you use your own data start here

################################################################################
################################################################################

#Load libraries
library(multilevel)

################################################################
#resample function:
#   creates ONE multi-stage bootstrap sample by
#   sampling with replacement at specific levels of
#   clustered data
#
# We followed:
#  http://biostat.mc.vanderbilt.edu/wiki/Main/HowToBootstrapCorrelatedData
#  but instead of dat[[cluster[1]]] in original function I wrote dat[,cluster[1]]
#  See explanations and documentation the wiki site
#
# This sampling technique is used in order to preserve the
#  nested correlation structure in resamples.
################################################################

#Arguments:
# dat     - data to bootstrap
#          We assume that the data has the following structure:
#          id group item1 item2 .... n_item
# cluster - A bivariate vector:
#          c("name identifying the clusters from which x originated",
#            "name of id variable") (see example below)
#
# replace - resampling scheme
#    The following resampling schemes are relevant for
#    hierarchical data:
#     - c(T,F) :Resampling only UPPER level,
#             but NOT resampling the LOWER level
#     - c(T,T) :Resampling BOTH UPPER and LOWER levels
#     - c(F.F) :NO resampling => original data
#     - c(F.T) :Resampling only LOWER level,
#             but NOT resampling the UPPER level

#Value
#   returns ONE bootstrap sample

resample <- function(dat, cluster, replace) {

  # exit early for trivial data
  if(nrow(dat) == 1 || all(replace==FALSE))
     return(dat)

  # sample the clustering factor
  cls <- sample(unique(dat[,cluster[1]]), replace=replace[1])

  #Instead of dat[[cluster[1]]] in original function I wrote dat[,cluster[1]]

  # subset on the sampled clustering factors
  sub <- lapply(cls, function(b) subset(dat, dat[,cluster[1]]==b))

  # sample lower levels of hierarchy (if any)
  if(length(cluster) > 1)
    sub <- lapply(sub, resample, cluster=cluster[-1], replace=replace[-1])

  # join and return samples
  do.call(rbind, sub)

}
#Example
Ex1_dat1<-read.csv("Ex1_dat1.csv") #read data
set.seed(123)
B1_Ex1_dat1 <- resample(dat = Ex1_dat1,
      cluster = c("group","id"),
      replace = c(T,T))#create 1 bootstrap sample

#print head of bootstrap sample
B1_Ex1_dat1[1:10,]
#print original data of the 1st bootstrapped group
Ex1_dat1[Ex1_dat1$group==B1_Ex1_dat1$group[1],]
```

```
#Example output:
#print head of bootstrap sample
#> B1_Ex1_dat1[1:10,]
#     id group item1 item2 item3 item4 item5 item6
#141  1    15     2     3     3     2     3     4
#145  5    15     4     2     3     2     2     2
#148  8    15     3     3     3     3     3     3
#142  2    15     2     2     2     1     2     1
#146  6    15     2     2     2     2     2     2
#143  3    15     3     2     3     2     3     2
#1421  2    15     2     2     2     1     2     1
#1481  8    15     3     3     3     3     3     3
#149  9    15     2     2     3     2     3     2
#144  4    15     1     1     1     2     1     1

#print original data of the 1st bootstrapped group
#> Ex1_dat1[Ex1_dat1$group==B1_Ex1_dat1$group[1],]
#    id group item1 item2 item3 item4 item5 item6
#141  1    15     2     3     3     2     3     4
#142  2    15     2     2     2     1     2     1
#143  3    15     3     2     3     2     3     2
#144  4    15     1     1     1     2     1     1
#145  5    15     4     2     3     2     2     2
#146  6    15     2     2     2     2     2     2
#147  7    15     3     3     3     3     3     3
#148  8    15     3     3     3     3     3     3
#149  9    15     2     2     3     2     3     2
#150 10    15     2     3     3     2     3     3

############################################################
#ci_med function:
#   calculate median and 95% percentile limits of vector x
#
#   95% percentile limits of vector x =
#      (2.5% percentile of x, 97.5% percentile of x)
#
#   The percentile limits will be denoted as CI
############################################################

#Arguments:
#x - vector
#Value
#   returns median and 95% percentile limits
ci_med<-function(x){
  cat("\nCI for",deparse(substitute(x)))
  cat("\nCI=(",round(quantile(x,probs=c(2.5,97.5)/100),3),")  median=",
      round(quantile(x,probs=c(50)/100),3),"\n")
 }
 #Example
x1<-c(0:1000) #create vector of numbers 0,1,2,...,1000
ci_med(x1)

#Example output:
#CI for x1
#CI=( 25 975 )  median= 500

##############################################################
#ci_med_C function:
#   calculate median and 95% percentile, mean, and percentile
#   limits of vector x
#
#   95% percentile limits of vector x =
#      (2.5% percentile of x, 97.5% percentile of x)
#
#   The percentile limits will be denoted as CI
#This function is similar to ci_med - but with shorter output
##############################################################

#Arguments:
#x - vector
#Value
#   returns median and 95% percentile limits

ci_med_C<-function(x){
  cat("\nCI=(",round(quantile(x,probs=c(2.5,97.5)/100),3),")  median=",
      round(quantile(x,probs=c(50)/100),3)," mean=",mean(x),"\n")
 }
 #Example
x1<-c(0:1000) #create vector of numbers 0,1,2,...,1000
ci_med_C(x1)

#Example output:
#CI=( 25 975 )  median= 500

########################################################
########################################################

#rwg_AD_ICC_sh function:
#  Prepares rwg AD ICC and their “matched difference”
#  calculations for a given data set.
#
#  rwg is calculated relative to reference probability
#  defined by pvec

########################################################

#Arguments:
#dat     - given data
#          We assume that the data has the following structure:
#          id group item1 item2 .... n_item
#n_item  - Number of items in data
#pvec    - Vector of reference probabilities

#Value
#   returns the following Outputs:

#   for rwg:
#   - rwg_A = mean of rwg.j values over the sample groups
#   - rwg_shuf = mean of rwg.j values over SHUFFLED sample groups
#   - rwg_dif  = rwg_A-rwg_shuf
#   - rel_rwg_dif = rwg_diff/rwg_shuf

#   for AD(The average deviation around the mean):
#   - AD_A = mean of AD values over the sample groups
#   - AD_shuf = mean of AD values over SHUFFLED sample groups
#   - AD_dif  = AD_A-AD_shuf
#   - rel_AD_dif = AD_diff/AD_shuf

#   for ICC:
# - ICC_Pv = ICC significance for the bootstrap sample
# - ICC_Pv_shuf = ICC significance for the SHUFFLED bootstrap sample
# - ICC1_A = ICC1 value for the bootstrap sample
# - ICC1_shuf = ICC1 value for the SHUFFLED bootstrap sample
# - ICC2_A = ICC2 value for the bootstrap sample
# - ICC2_shuf  = ICC2 value for the SHUFFLED bootstrap sample
# - dif_ICC1 = ICC1_A - ICC1_shuf
# - dif_ICC2 = ICC2_A - ICC2_shuf

rwg_AD_ICC_sh <- function(dat, n_item, pvec) {
  a<-length(pvec)

  #Shuffle data
  dat_shuf<-dat[,3:(2+n_item)][sample(nrow(dat)),]

  #Rwg.j
  rwg_A<-mean(rwg.j( dat[,3:(2+n_item)],
             grpid =  as.factor(dat$group),
             ranvar=(a**2 - 1)/12)$rwg.j )
```

```
 rwg_shuf<-mean(rwg.j( dat_shuf,grpid =  as.factor(dat$group),
           ranvar=(a**2 - 1)/12)$rwg.j)
 rwg_dif=rwg_A-rwg_shuf
 rel_rwg_dif=rwg_dif/rwg_shuf

 #ad.m
 AD_A<-mean(ad.m( dat[,3:(2+n_item)],
          grpid =  as.factor(dat$group),
          type="mean")$AD.M)
 AD_shuf<-mean(ad.m( dat_shuf,grpid =  as.factor(dat$group),
           type="mean")$AD.M)
 AD_dif=AD_A-AD_shuf
 rel_AD_dif=AD_dif/AD_shuf

 #ICC
 dep<-apply(dat[,3:(2+n_item)],1,mean)
 hrs.mod<-aov(dep~as.factor(dat$group))

 dep_shuf<-apply(dat_shuf,1,mean)
 hrs.mod_shuf<-aov(dep_shuf~as.factor(dat$group))

 ICC_Pv<-summary(hrs.mod)[[1]][1,c("Pr(>F)")]
 ICC_Pv_shuf<-summary(hrs.mod_shuf)[[1]][1,c("Pr(>F)")]

 ICC1_A<-ICC1(hrs.mod)
 ICC1_shuf<-ICC1(hrs.mod_shuf)

 ICC2_A<-ICC2(hrs.mod)
 ICC2_shuf<-ICC2(hrs.mod_shuf)

 dif_ICC1= ICC1_A - ICC1_shuf
 dif_ICC2= ICC2_A - ICC2_shuf

  c(rwg_A,rwg_shuf,rwg_dif,rel_rwg_dif,
   AD_A,AD_shuf,AD_dif,rel_AD_dif,
   ICC_Pv,ICC_Pv_shuf,
   ICC1_A,ICC1_shuf,ICC2_A,ICC2_shuf,dif_ICC1,dif_ICC2)
}
#Example
Ex1_dat1<-read.csv("Ex1_dat1.csv") #read data
set.seed(881)
out_Ex1 <- rwg_AD_ICC_sh(Ex1_dat1, n_item=6, pvec=c(.2,.2,.2,.2,.2))
names(out_Ex1)<-c("rwg_A","rwg_shuf","rwg_dif","rel_rwg_dif",
           "AD_A", "AD_shuf" ,"AD_dif", "rel_AD_dif",
           "ICC_Pv","ICC_Pv_shuf",
           "ICC1_A","ICC1_shuf","ICC2_A",
            "ICC2_shuf","dif_ICC1","dif_ICC2")
out_Ex1

#Example output:
#> out_Ex1
#      rwg_A     rwg_shuf     rwg_dif  rel_rwg_dif        AD_A      AD_shuf
# 0.769248681  0.805846092 -0.036597410 -0.045414888  0.754533333  0.779666667
#     AD_dif   rel_AD_dif      ICC_Pv  ICC_Pv_shuf      ICC1_A    ICC1_shuf
#-0.025133333 -0.032235998  0.014067805  0.309599704  0.051014406  0.009609624
#     ICC2_A    ICC2_shuf     dif_ICC1     dif_ICC2
# 0.349622157  0.088446779  0.041404781  0.261175377

########################################################
# Create multiple 2 levels Bootstrap samples
#       with extended output
########################################################

#Note:
#In order to create bootstrap CI RUN:
# - rep_boot:to create n bootstrap samples and calculate RWG AD ICC
#           and their “matched difference” estimates for each
#           bootstrap sample
#
#           rep_boot function calls f_dat_repl which creates ONE
#           bootstrap samples and calculate RWG AD ICC
#           and their “matched difference” estimates for this
#           bootstrap sample
#
# - ci_ex:   to create the CI and point estimates out of these estimates
#See example below

##############
# f_dat_repl
##############

#function f_dat_repl: create 1 bootstrap sample
#                   and keep statistics for them

#Arguments:
# Ex_dat - data to bootstrap
#         We assume that the data has the following structure:
#         id group item1 item2 .... n_item
# n_item - Number of items in data
# pvec   - Vector of reference probabilities
# replace - resampling scheme
#     The following resampling schemes are relevant for
#     hierarchical data:
#      - c(T,F) :Resampling only UPPER level,
#             but NOT resampling the LOWER level
#      - c(T,T) :Resampling BOTH UPPER and LOWER levels
#      - c(F.F) :NO resampling => original data
#      - c(F.T) :Resampling only LOWER level,
#             but NOT resampling the UPPER level

f_dat_repl <- function(Ex_dat,nitems, pvec,replace) {
  cluster <- c("group","id")

  #Create one bootstrap sample
  dat<-cbind(Ex_dat[,1:2],resample(Ex_dat, cluster,replace )[,3:(2+nitems)])

  #Create RWG and ICC statistics
  rwg_AD_ICC_sh(dat, nitems, pvec)
}
##############
# rep_boot
##############

#function rep_boot: create n bootstrap samples and
#                keep statistics for them
#Arguments:
# n - number of bootstrap replicates
# Ex_dat - data to bootstrap
#         We assume that the data has the following structure:
#         id group item1 item2 .... n_item
# n_item - Number of items in data
# pvec   - Vector of reference probabilities
# replace - resampling scheme
#     The following resampling schemes are relevant for
#     hierarchical data:
#      - c(T,F) :Resampling only UPPER level,
#             but NOT resampling the LOWER level
#      - c(T,T) :Resampling BOTH UPPER and LOWER levels
#      - c(F.F) :NO resampling => original data
#      - c(F.T) :Resampling only LOWER level,
#             but NOT resampling the UPPER level

rep_boot <- function(n,Ex_dat,nitems, pvec,replace){
  res<-replicate(n, f_dat_repl(Ex_dat,nitems, pvec,replace))
  tres<-as.data.frame(t(res))
  names(tres)<-c("rwg_A","rwg_shuf","rwg_dif","rel_rwg_dif",
             "AD_A", "AD_shuf" ,"AD_dif", "rel_AD_dif",
             "ICC_Pv","ICC_Pv_shuf",
              "ICC1_A","ICC1_shuf","ICC2_A",
              "ICC2_shuf","dif_ICC1","dif_ICC2")
  tres
}
##############
# ci_ex
##############

#function ci_ex: Calculate bootstrap CI for  RWG AD ICC and their
#               “matched difference” statistics
#               stored in boot_res
#Arguments:
# boot_res - object created by rep_boot function

ci_ex <- function(boot_res){

 cat("\n== RWG ==")
 ci_med_C(boot_res$rwg_A)

 cat("\n== RWG  shuffled ==")
 ci_med_C(boot_res$rwg_shuf)

 cat("\n== RWG  dif ==")
 ci_med_C(boot_res$rwg_dif)

 cat("\n== rel_rwg_dif:relative RWG  dif ==")
 ci_med_C(boot_res$rel_rwg_dif)

 cat("\n== AD ==")
 ci_med_C(boot_res$AD_A)

 cat("\n== AD shuffled ==")
 ci_med_C(boot_res$AD_shuf)

 cat("\n== AD  dif ==")
 ci_med_C(boot_res$AD_dif)

 cat("\n== rel_AD_dif:relative AD  dif ==")
 ci_med_C(boot_res$rel_AD_dif)

 cat("\n== ICC1 ==")
 ci_med(boot_res$ICC1_A)

 cat("\n== ICC1 shuffled ==")
 ci_med(boot_res$ICC1_shuf)
 cat("\n== ICC1  dif ==")
 ci_med(boot_res$dif_ICC1)

 cat("\n== ICC2 ==")
 ci_med(boot_res$ICC2_A)

 cat("\n== ICC2 shuffled ==")
 ci_med(boot_res$ICC2_shuf)

 cat("\n== ICC2  dif ==")
 ci_med(boot_res$dif_ICC2)
}
#Example
Ex1_dat1<-read.csv("Ex1_dat1.csv") #read data
set.seed(881)

boot_Ex1_dat_1<-rep_boot(n = 1000,Ex_dat = Ex1_dat1,nitems = 6,
    pvec = c(.2,.2,.2,.2,.2),replace = c(T,F)) #bootstrap

ci_ex(boot_Ex1_dat_1) #calculate bootstrap CI

#Example output:
#== RWG ==
#CI=( 0.694 0.835 )  median= 0.77  mean= 0.7690101
#
#== RWG  shuffled ==
#CI=( 0.714 0.863 )  median= 0.803  mean= 0.8010393
#
#== RWG  dif ==
#CI=( -0.073 0.009 )  median= -0.032  mean= -0.03202914
#
#== rel_rwg_dif:relative RWG  dif ==
#CI=( -0.091 0.012 )  median= -0.04  mean= -0.03967611
#
#== AD ==
#CI=( 0.69 0.829 )  median= 0.755  mean= 0.755894
#
#== AD shuffled ==
#CI=( 0.71 0.855 )  median= 0.779  mean= 0.7795229
#
#== AD  dif ==
#CI=( -0.053 0 )  median= -0.023  mean= -0.02362887
#
#== rel_AD_dif:relative AD  dif ==
#CI=( -0.065 0.001 )  median= -0.029  mean= -0.03012699
#
```

```
#== ICC1 ==
#CI for boot_res$ICC1_A
#CI=( -0.015 0.115 )  median= 0.047
#
#== ICC1 shuffled ==
#CI for boot_res$ICC1_shuf
#CI=( -0.038 0.044 )  median= -0.002
#
#== ICC1  dif ==
#CI for boot_res$dif_ICC1
#CI=( -0.029 0.124 )  median= 0.048
#
#== ICC2 ==
#CI for boot_res$ICC2_A
#CI=( -0.179 0.565 )  median= 0.332
#
#== ICC2 shuffled ==
#CI for boot_res$ICC2_shuf
#CI=( -0.582 0.315 )  median= -0.015
#
#== ICC2  dif ==
#CI for boot_res$dif_ICC2
#CI=( -0.246 0.977 )  median= 0.341
```