2007

Does the Measure of Dispersion Matter in Multilevel Research? A Comparison of the Relative Performance of Dispersion Indexes

Quinetta M. Roberson
Cornell University

Michael C. Sturman
Cornell University, mcs5@cornell.edu

Tony L. Simons
Cornell University, tls11@cornell.edu

Recommended Citation

Roberson, Q. M., Sturman, M. C., & Simons, T. L. (2007). *Does the measure of dispersion matter in multilevel research? A comparison of the relative performance of dispersion indexes* [Electronic version]. Retrieved [insert date], from Cornell University, School of Hospitality Administration site: http://scholarship.sha.cornell.edu/articles/101

Does the Measure of Dispersion Matter in Multilevel Research? A Comparison of the Relative Performance of Dispersion Indexes

Abstract

Within the context of climate strength, this simulation study examines the validity of various dispersion indexes for detecting meaningful relationships between variability in group member perceptions and outcome variables. We used the simulation to model both individual-and group-level phenomena, vary appropriate population characteristics, and test the proclivity of standard and average deviation, interrater agreement indexes (r_{wg} , r^*_{wg} , a_{wg}), and coefficient of variation (both normed and unnormed) for Type I and Type II errors. The results show that the coefficient of variation was less likely to detect interaction effects although it outperformed other measures when detecting level effects. Standard deviation was shown to be inferior to other indexes when no level effect is present although it may be an effective measure of dispersion when modeling strength or interaction effects. The implications for future research, in which dispersion is a critical component of the theoretical model, are discussed.

Keywords

multilevel research, climate strength, dispersion indexes, level effects, perceptions, outcome variables

Disciplines

Statistical Theory

Does the Measure of Dispersion Matter in Multilevel Research?

A Comparison of the Relative Performance of Dispersion Indices

Quinetta M. Roberson
Human Resource Studies
Cornell University
393 Ives Hall
Ithaca, NY 14853-3901
Phone: 607-255-4454
Fax: 607-255-1836
Email: QMR3@cornell.edu

Michael C. Sturman
School of Hotel Administration
Cornell University
545F Statler Hall
Ithaca, NY 14853
Phone: 607-255-5383
Fax: 607-254-2971
Email: MCS5@cornell.edu

Tony L. Simons
School of Hotel Administration
Cornell University
538 Statler Hall
Ithaca, NY 14853
Phone: 607-255-8382
Fax: 607-254-2971
Email: TLS11@cornell.edu

Abstract

Within the context of climate strength, we conduct a simulation study to examine the validity of various dispersion indices for detecting meaningful relationships between variability in group member perceptions and outcome variables. We used the simulation to model both individual- and group-level phenomena, vary appropriate population characteristics, and test the proclivity of standard and average deviation, interrater agreement indexes (r_{wg}, $r*_{wg}$, a_{wg}), and coefficient of variation (both normed and unnormed) for Type I and Type II errors. The results showed that the coefficient of variation was less likely to detect interaction effects although it outperformed other measures when detecting level effects. Standard deviation was shown to be inferior to other indices when no level effect is present although it may be an effective measure of dispersion when modeling strength or interaction effects. The implications for future research, in which dispersion is a critical component of the theoretical model, are discussed.

As researchers increasingly adopt a multilevel approach to better understand organizational phenomena, attention has been given to issues of multilevel construct validation. More specifically, researchers have articulated relevant composition processes, or functional relationships among constructs at different levels that have the same content, meaning, and nomological network but are qualitatively different (Kozlowski & Klein, 2000). Although several composition models have been proposed (see Chan, 1998), researchers have primarily given attention to consensus models, which use within-group agreement of individual responses to demonstrate structural and functional equivalence between a lower-level construct and a higher-level construct (Morgeson & Hoffman, 1999). Because consensus is a necessary condition for construct validity at the higher level, indices of agreement are used to determine whether there is sufficient consensus at the lower level to justify aggregation of individual responses to represent the higher-level construct (e.g., James, Demaree & Wolf, 1993; Kozlowski & Hattrup, 1992).

Researchers have recently begun considering compilation models of emergence, which describe constructs at different levels that are similar in function but distinctly different in meaning and form (Morgeson & Hoffman, 1999). One fundamental assumption of compilation models is that because the kinds of contributions that individuals make to the group are dissimilar, member responses will not necessarily converge (Kozlowski & Klein, 2000). Given a lack of structural equivalence between lower-level and higher-level constructs, high agreement is not a prerequisite for establishing construct validity at the higher level (Bliese, 2000). Instead, within-group variance in individual responses is treated as a focal construct rather than a statistical prerequisite for aggregation (Kozlowski & Klein, 2000). Therefore, dispersion indices are used to capture the variability among individual characteristics, responses, or contributions to the group (Lindell & Brandt, 1999). Further, because the amount of within-group variance is expected to also vary across groups and be associated with important outcomes, such indices have been used as independent variables in subsequent analyses (e.g., Colquitt, Noe & Jackson, 2002; Gonzalez-Roma, Peiro & Tordera, 2002; Lindell & Brandt, 2000; Schneider, Salvaggio & Subirats, 2002).

Although researchers have compared measures of within-group agreement (Brown & Hauenstein, 2005; Burke & Dunlap, 2002; Burke, Finkelstein & Dusig, 1999; Kozlowski & Hattrup, 1992), there has been little consideration given to the quality and efficacy of such indices as measures of dispersion. Further, while researchers have distinguished between different dispersion indices based on the theoretical and methodological issues associated with each index, they have not done so within the context of compilation models of emergence. Some research has demonstrated that indices of agreement are highly correlated (Burke et al., 1999). However, such research also highlights the proportion of variance that is not shared by the indices. For example, while Burke et al. (1999) showed correlations between average deviation and interrater agreement indices (r_{wg}; James et al., 1993) to be between 0.79 and 0.93, such results also suggest that 14% – 38% of their variance is not shared. Consequently, measures of dispersion may yield different inferences regarding the relationship between within-group variance and group-level outcome variables. It may also be the case that the relationships between various dispersion indices are nonlinear, such that correlations may not fully capture the nature of the relationships between indices. Therefore, an inclusive comparison of dispersion measures is needed.

Within the context of psychological climates, we conduct a simulation study to examine the validity of various dispersion indices for detecting meaningful relationships between group member perceptions and outcome variables. We used the simulation to model both individual- and group-level phenomena, vary appropriate population characteristics, and test the proclivity of the various measures for both Type I and Type II errors. We also examined the efficacy of these measures for modeling main effects and interactions, both with and without specification error. The implications for future research, in which dispersion is a critical component of the theoretical model, are discussed.

Dispersion Measures

Although dispersion models are theoretically appropriate for a variety of constructs, many organizational researchers have examined dispersion composition within the context of climate perceptions, or employee perceptions of policies, practices and procedures that characterize an organization or work unit (Schneider & Reichers, 1983). Specifically, the effects of climate strength, or variability in group member climate perceptions (Lindell & Brandt, 2000; Schneider et al., 2002), has been examined. For example, Lindell and Brandt (2000) explored the mediating effects of organizational climate strength in the relationship between climate quality and organization-level outcomes. Similarly, Colquitt and his colleagues (2002) considered the direct effects of team justice climate strength on team outcomes, such as performance and absenteeism. The results of these studies, however, did not show significant direct effects of climate strength.

Researchers have also examined climate strength as a moderator in the relationships between climate level (i.e., mean climate perceptions) and unit-level outcomes (Colquitt et al., 2002; Gonzalez-Roma et al., 2002; Schneider et al., 2002). Based on the concept of situational strength (Mischel, 1973), which refers to the extent to which a context is unstructured and ambiguous and thus guides social behavior, researchers have proposed that climate strength affects the predictability of individual responses (Lindell & Brandt, 2000). For example, because strong situations are those in which people have similar interpretations of events and uniform expectations about appropriate behavior, variability in individuals' responses to such situations will be low (Mischel, 1973). Consequently, behavioral predictions should be more reliable in structured and unambiguous situations, such as strong climates. In contrast, weak climates should produce less salient cues for interpreting events and guiding behavior, thus reducing the consistency in, and predictability of, individual behaviors. Some research provides evidence that climate strength may facilitate the influence of climate level perceptions on group-level outcomes (Colquitt et al., 2002; Gonzalez-Roma et al., 2002).

Even with a limited amount of research on the predictive effects of climate strength, perceptual dispersion has been conceptualized and measured in different ways. For example, Lindell and Brandt (2000) suggest that an estimate of interrater agreement – e.g., the r_{wg} index of interrater reliability (James et al., 1984) – can appropriately serve as a dispersion index. Alternatively, researchers have relied upon diversity indices, such as the coefficient of variation, which corrects for the lack of independence between measures of central tendency and measures of dispersion, to indicate perceptual variability (see Colquitt et al., 2002). Researchers have also suggested that dispersion measures, such as standard deviation or an average deviation index, might be more appropriate representations of variation in group members' responses on a given measure (Bedeian & Mossholder, 2000; Lindell & Brandt, 2000). To develop a framework for comparing indices of climate strength, we provide an overview of these dispersion indices.

Standard Deviation

Standard deviation is considered to be an appropriate index for representing a lack of consensus or agreement within a focal population (Schmidt & Hunter, 1989). Calculated as the average (squared) distance of a set of scores from the mean, standard deviation measures the spread of data around the mean. Research has shown that standard deviation statistics are useful for both binary and non-binary response scales and can be used across any population with two or more individual scores on the same measure (Conway & Schaller, 1998). In addition, standard deviation is relatively easy to calculate and to understand relative to other measures of dispersion (see Table 1).

Despite these operational benefits, researchers have noted several problems with using standard deviation as a measure of dispersion (see Kozlowski & Hattrup, 1992). Because the upper boundary on the measurement scale is determined by the values on the response scale, it is difficult to establish the maximum possible value of the standard deviation statistic. Actual values on the measurement scale are scale-specific, which limits the comparison of standard deviation statistics from different empirical investigations. Therefore, the interpretation of the standard deviation statistic is limited by the lack of clear anchor points on measurement scales. Standard deviation as a measure of dispersion is also limited by its sensitivity to anomalous values or longer-tailed distributions. By squaring the difference between a value and the mean, the standard deviation potentially provides a distorted view of the amount of dispersion in a set of values. More specifically, the act of squaring makes each unit of distance from the mean exponentially (rather than additively) greater, which is not completely eliminated by calculating the square-root of the sum of squares. As such, the influence of extreme values is increased.

Another limitation of standard deviation measures is that they do not lend themselves easily to inferential statistical inquiries given that values may differ based on whether standard deviation is calculated from all scores in a population of interest or from scores in a sample from that population (Conway & Schaller, 1998). The key difference between the formulae is that sample standard deviation uses (n-1) in the denominator and is therefore, larger than the population standard deviation which uses N. Although this difference may be immaterial in studies in which group sizes are consistent, it may negatively influence the inferential interpretation of effects in studies in which group sizes differ. Because sample standard deviation is influenced by the actual degree of dispersion within a group of individual scores as well as the number of individual scores within that group, the use of a sample formula may introduce error variance and decrease statistical power relevant to inferential judgment under circumstances in which group size is a random variable of no conceptual importance. Alternatively, if group size is an a priori variable of conceptual interest, the use of sample standard deviation may confound size with method variance. Consequently, the standard deviation calculation may lead researchers to make Type I errors.

Average Deviation

Although there are several alternatives to standard deviation as a measure of dispersion, the most direct alternative is the absolute mean deviation, which is also referred to as the average deviation index. Calculated as the average of the absolute differences between each score and the overall mean (or median), the average deviation index is considered to be a more useful measure of dispersion than standard deviation given that it is interpretable in terms of the metric of the original scale (Burke et al., 1999). Research has also shown that the average deviation index is better for use with distributions other than a normal distribution given that it is less sensitive to extreme scores or deviations from normality (Stigler, 1973). Further, unlike variance ratio indices (e.g., r_{wg}), the use of an average deviation index does not require explicitly modeling the random or null response distribution.

Despite several benefits of using average deviation rather than standard deviation, researchers have cautioned against the possibility of non-random response bias as well as unrepresentative values for rater responses when the data include extreme values (Burke et al., 1999). In addition, the average deviation index has been typically used as a measure of agreement although the index itself is actually a measure of variability, given that a score of zero would be indicative of complete agreement among raters (Burke & Dunlap, 2002; Burke et al., 1999). Therefore, research is needed to apply the average deviation index to various situations and content domains to evaluate its efficacy as a measure of dispersion.

Interrater Agreement Indices (r_{wg}, $r*_{wg}$, and a_{wg})

Developed by James, Demaree and Wolf (1984), r_{wg} compares the variability of a given variable within a specific unit to an expected variance. Computed using either an individual item ($r_{wg(I)}$) or multiple items ($r_{wg(J)}$)[1], the resulting score estimates the degree to which observed similarity in responses is due to actual agreement between unit members. Unlike some indices that assess internal agreement (or a lack thereof) across a set of potential aggregation units, r_{wg} is calculated separately for each unit. This is considered to be a key strength of the r_{wg} index given that agreement is not based on between-group variability (James et al., 1993). Further, because r_{wg} values fall between 0 and 1, the interpretation is considered to be relatively straightforward (Conway & Schaller, 1998).

There are several assumptions associated with the r_{wg} index. Specifically, the measure is intended to be used in analyzing variables that have a unidimensional factor structure (James et al., 1984), discrete response formats (Castro, 2002), and nearly equal measurement intervals (James et al., 1984). In the calculation of r_{wg}, expected variance is usually operationalized as the variance of a uniform (i.e., rectangular) distribution. Accordingly, the use of r_{wg} requires empirical evidence that supports the null distribution and that there is one true score underlying individual responses, although there is no true score variance when a single stimulus is rated (Schmidt & Hunter, 1989). The use of r_{wg} also requires evidence that the response distribution is not bimodal or multimodal (LeBreton, James & Lindell, 2005). However, because psychological responses are subject to response bias and therefore non-random, the r_{wg} index is typically overstated (Klein & Kozlowski, 2000). Conditions under which unit members' responses are polarized, or at the extremes of the response scale, are also problematic given that the null distribution will result in an understatement of r_{wg} (Lindell, Brandt & Whitney, 1999).

Another limitation of r_{wg} is that, like other variance measures, it is affected by sample size. Specifically, the interpretation of low agreement values can be difficult in situations where sample size is small (Kozlowski & Hattrup, 1992). The comparison of r_{wg} indices across studies may also be complicated given that such measures are dependent upon expected variance, which may differ across samples (Conway & Schaller, 1998), and the number of rating scale anchors, which may change the lower bound of the index across studies (Brown & Hauenstein, 2005). Research suggests that the magnitude of $r_{wg(J)}$ values may also be influenced by the number of items in a measure, which may subsequently affect the probability of obtaining values greater than zero (Schriesheim, Cogliser & Neider, 1995). Beyond issues of interpretation, researchers have noted the difficulty in conducting inferential statistical inquiries using r_{wg} as a dispersion measure. Specifically, because significance tests of r_{wg} are dependent upon sample size, number of items in a measure, and number of rating scale anchors, computer-intensive methods, such as bootstrapping or Monte Carlo simulations, must be used to estimate confidence intervals and compare variances between independent groups (Cohen, Dovey & Eick, 2001; Conway & Schaller, 1998).

Lindell and his colleagues (1999) proposed a modified index of interrater agreement, r^*_{wg}, to address some of these limitations. Because it is possible for observed variances to exceed the variance of the uniform distribution in the calculation $r_{wg(J)}$, the index may not accurately assess differences in the observed variances as the number of items in the scale increases. Consequently, researchers using the $r_{wg(J)}$ index to examine multi-item rating scales with large numbers of items may detect a sufficient amount of agreement in a set of ratings when such agreement does not exist. Lindell et al. (1999) propose that r^*_{wg}, which is an inverse linear function of the ratio of the average observed variance to the variance of uniformly distributed random error, avoids such inaccuracies. By substituting the average item variance for the observed variance in respondents ratings, $r^*_{wg(J)}$, which is used for multi-item scales, is not dependent upon the number of items in a scale. In addition, the index allows for better interpretation by equating zero with random response and representing less than hypothesized levels of observed agreement by negative values. Accordingly, r^*_{wg} can be used as an index of agreement (see Lindell & Brandt, 2000), or disagreement.

Brown and Hauenstein (2005) also propose an alternative interrater agreement index (a_{wg}) to overcome the limitations of r_{wg}. Based on the principles of Cohen's (1960) kappa statistic, which is used to assess agreement between judges rating multiple stimuli on a categorical scale, $a_{wg(J)}$ is estimated using multiple null distributions rather than one specification of the null distribution. By using the maximum possible variance at the mean as the null distribution, a_{wg} is interpretable as the proportion of consensus to maximum possible disagreement. Therefore, unlike the interrater agreement index proposed by James et al. (1984), a_{wg} values are not dependent upon sample size, scale, or the location of the observed means. Brown and Hauenstein (2005) recognize that because a_{wg} is computed with the observed mean and observed variance of ratings, the measure is susceptible to sampling error and may be influenced by the number of raters. However, given that values of a_{wg} range from -1 to +1, indicating maximum disagreement to absolute agreement, the interpretability of the index along with its other advantages may outweigh the aforementioned limitations.

Coefficient of Variation

The coefficient of variation (often expressed as V) was derived to compare the variability of multiple potential aggregation units that have widely differing means. Calculated by dividing a sample's standard deviation by its mean, it indicates within-group differences among scores on a response variable in comparison to their average magnitude. Because the numerator and denominator are expressed in the same units, the coefficient of variation is scale-independent and is therefore considered useful for indicating variability across samples in relative terms (Allison, 1978). Further, the coefficient of variation is intended to be an improvement over other measures given that it represents the dispersion of a dataset relative to its own mean, which serves to reduce the influence of absolute size on variability (Bedeian & Mossholder, 2000).

In organizational research, the coefficient of variation has been used to describe within-group variability on a variety of perceptual and attitudinal variables. For example, Colquitt and his colleagues (2002) used the coefficient of variation to represent justice climate strength, or the variation (or lack thereof) in team members' ratings of justice. Other researchers, however, have argued against the use of this measure of dispersion with variables measured on an interval scale given that such variables do not have true fixed zero-points (Allison, 1978; Bedeian & Mossholder, 2000). More specifically, they assert that sample means and standard deviations are arbitrarily derived given that most interval-level variables can be characterized by arbitrary zero points and ranges. The use of the coefficient of variation also assumes the existence of a nonnegative ratio scale underlying the interval scale (Allison, 1978), which may not be the case for many psychological variables – in particular, climate perceptions. The utility of the coefficient of variation may be further limited by its calculation, based on both a measure of dispersion and the group mean, which constrains the opportunity for meaningful comparisons across samples. Because it is influenced by differences in group or sample sizes (Martin & Gray, 1971), Bedeian and Mossholder (2000) suggest that researchers should adjust for such differences when using the coefficient of variation. Specifically, they argue that a normalized coefficient of variation (often expressed as V", as proposed by Smithson, 1982), may better represent equivalence among groups that differ in sample size.

Distinguishing Between Measures of Dispersion

In comparative analyses of agreement indices, researchers have used a priori decision rules to determine how well each measure captures interrater agreement (see Brown & Hauenstein, 2005; Burke et al., 1999; Kozlowski & Hattrup, 1992). Specifically, indices of agreement have been compared based on the probability of reaching an agreement threshold as well as ease of interpretation. Based on the results of such analyses, which highlight the methodological limitations discussed above, researchers have concluded that decisions about consensus are influenced by the choice of agreement index and therefore, have supported the use of specific agreement statistics. From a dispersion perspective, however, different conclusions about the validity of these measures may be reached. Under compilation models of emergence, within-group variance is treated as a central construct rather than as a statistical requirement for aggregation (Chan, 1998; Klein & Kozlowski, 2000). As such, dispersion statistics are used to represent the distribution of individual responses within a group rather than evidence of the shared properties of the group. Although prior research highlights the usefulness of various indices to indicate the level of consensus or dissensus within a group, we have little understanding of how such indices influence inferences about group-level relationships in a dispersion model. Beyond the probability of Type I and Type II errors, a number of questions also remain about how changes in sample characteristics or parameter estimates influence the predictive power of dispersion indices. In the following section, we conduct a simulation address these issues.

Methods

Simulation Design and Parameters

We used computer simulation to develop a set of circumstances with which to compare standard deviation, average deviation index, r_{wg}, $r*_{wg}$, a_{wg}, and coefficient variation (both unnormed and normed). The key first step in creating such a simulation is to create a model of reality that is feasible to simulate but also provides results that are useful for evaluating the specific research question. For our purposes, we simulated a situation where (1) there is an unobservable group-level construct, (2) individuals perceive that construct to varying degrees, resulting in individual-level observations of the variable that are the source for surrogating the group-level construct, and (3) the group-level construct and within-group variance in the individual measures affect a second group-level variable that is observed. This representation of the universe for the simulation is shown in Figure 1.

The measurement model for our simulation is presented in Figure 2. This figure depicts a situation in which a researcher might try to approximate a group-level outcome (e.g., team performance) by examining the effects of a group-level construct derived from observed individual-level measures, the total amount of individual variance explained by the group construct, and their interaction. For each group, we simulated individual observations ranging from 1 to E, the number of employees per group. These observations represented individual-level measures of the group-level construct, as indicated in the figure. Once a set of observations was simulated for each group, the data was used to compute measures of climate level, calculated as the mean of the E observations per group, and climate strength, calculated using one of the aforementioned dispersion measures. Regression analysis was then used to examine the influence of climate level (B_1), climate strength (B_2), and the interaction of climate level and strength (B_3) on the group-level outcome. These procedures allowed us to examine how the operationalization of dispersion affects (1) the likelihood of Type I errors, and (2) the ability to detect significant relationships between group member perceptions and outcome variables.

A summary of the study's parameters, which were used to provide a broad array of realistic conditions, and their levels are shown in Table 2. We varied simulation parameters to represent values that we might expect to observe in organizational research, or used values that reflected current practices in the field. To represent a realistic set of samples, we varied both group size (number of employees per group) and number of groups in the analyses. Prior justice climate research has been conducted with group sizes ranging from 3 – 90 members and number of groups ranging from approximately 40 – 250 groups (see Colquitt et al., 2005 for a review). However, given the limited number of studies examining justice at the group level of analysis, we examined these sample characteristics in multilevel research published from 2000 – 2005.[2] Because some researchers have considered groups to consist of as few as three members, we used three as a minimum group size. In addition, we considered five and ten to be reasonably larger yet common group sizes, and 25 members to represent larger groups. To model the range of sample sizes (in terms of number of groups) included in prior research, we simulated situations with 40, 80, or 120 groups.

Relationship between group-level construct and individual-level measures

To represent a variety of potential relationships between the group-level construct and individual-level approximations of the construct, we varied the percent of variance in the individual-level measures explained by the group-level construct and the variability of this relationship. The resultant total variability – that is, the total amount of individual variance explained by the group construct – is the representation of climate strength. Because this is a simulation, we were able to specify the amount of individual-level variance explained by group membership. This is the method used in other simulations with individual-level and group-level data (e.g., Bliese, 1998). As shown in table 2, the base amount of individual variance explained by the group construct was either none (0%), small (10%), medium (30%), or large (50%).

We also needed to contrast the extent to which this level of variance differed across groups. As noted in our introduction, the purpose of examining variability across groups is because this variability, as a unit-level construct, may be a predictor of unit-level outcomes. Accordingly, we needed to consider cases where variability was none, small (increasing the percent of variance explained by the group construct up to 10 percentage points), medium (increasing the percent of variance explained by the group construct up to 30 percentage points) and large (increasing the percent of variance explained by the group construct up to 50 percentage points). We chose the amount of variability (i.e., up to 10%, 30%, or 50%) to be from a uniform distribution, so that all values in the range were equally likely.

As shown in Figure 1, the total amount of individual variance explained by the group construct (δ_T), or climate strength, is equal to the base amount (δ_B) plus a random number from a continuous distribution, of 0 to δ_V. Therefore, the amount of the individual-level variance explained by group membership across all of our simulation scenarios ranged from 0% (no within-group variance) to 100% (50% within-group variance plus up to 50% potential within-group variance under the condition of maximal variability). Once the individual-level variables were created, we converted the individual-level scores to be on a seven-point scale, with a mean of 4.0, a standard deviation of 1.4, and truncated at 1 and 7. This conversion made the distribution of individual-level scores comparable to a Likert-type scale from 1 to 7.

Relationship between the group-level predictors and group-level outcome

Once the data for all individuals and each group were created, the simulation then generated the group-level outcome, using a combination of simulation parameters and the group-level constructs created in the simulation. As shown in Figure 1, the group-level outcome is a function of the effects of group-level construct, the total amount of individual variance explained by the group construct, and their interaction. For each of these effects, we used standardized regression coefficients. Based on Cohen (1992), which describes the magnitude of effect sizes, we chose standardized coefficients to represent no effect (0.00), small effects (0.10), medium effects (0.30), and large effects (0.50). Error was added in proper proportion so that the percent variance explained by each coefficient was β-squared.

Because the value of the group-level construct (i.e., climate level) was already expressed as a *Z*-score, it was simply multiplied by β_1 to represent its effects on the group outcome variable. However, the effect of the total amount of individual variance explained by the group construct (i.e., climate strength) was more complex to model. While the true percent of individual-level variance explained by the group-level construct is known through the simulation (i.e., δ_T), the values are not standardized, and thus cannot simply be multiplied by the desired beta coefficient to create the appropriate amount of variance explained in the dependent variable. For the purposes of standardizing the variable, we ran the simulation with 10 cases per scenario (details on running the simulation are provided below) and saved the value of one minus the true percent variance explained for each group within each simulation. The reason we subtracted the true percent variance explained from one was so that higher numbers represented greater dispersion (i.e., “0” represented no dispersion within a group while “1” represented the highest level of dispersion within a group). This resulted in 7,987,200 values, with a mean of 0.6714 and standard deviation of 0.1967. We used this calculated mean and standard deviation to convert climate strength values into standardized units before multiplying it by the beta coefficient (β_2) in the simulation.

The interaction effect was modeled by multiplying the group-level construct (i.e., climate level) with the standardized measure of the total amount of individual variance explained by the group construct (i.e., climate strength). This product was multiplied by the specified beta coefficient (β_3) and added to the previously described products. Finally, random error was added to yield the final value of the group-level outcome variable.

Simulation Implementation

The simulation was written in Visual Basic (Microsoft, 1998). Code for the simulation program was taken from DataSim (Sturman, 2004); however, we used a customized program for the purposes of generating the data for this study (a data generation algorithm is included in the appendix). The simulation generates two text files – one with the scenario-based results (i.e., the regression results) and one with the group-based results (i.e., data on each group within each scenario and the seven measures of dispersion).

Overall, the simulation included seven parameters, six parameters with four levels and one parameter with three levels. In all, this could produce a total of 12,288 different combinations of parameters. We did not, however, use a completely factorial design. While we were interested in varying both the magnitude of the climate strength effect and the magnitude of the variability of variance for the group-to-individual relationship, it did not make sense to simulate a climate strength effect when variability was zero. That is, if variability was zero, we would not expect to capture any relationship between the dispersion measures and the outcome (hence, significance would be a Type I error); at the same time, if a climate strength effect was present, then the dispersion measures should be related to the outcome (hence, a lack of significance would be a Type II error). Because of this contradiction, these cases were removed from consideration (i.e., the cases where the variability of variance of the group-to-individual relationship is zero—1/4 of the simulations— and where the variability of climate strength effect was small, medium, or large—3/4 of the simulations). Thus, 3/16 of cases were removed, leaving a total of 9,984 different scenarios.

We ran 100 cases for each scenario, resulting in a total of 998,400 cells of data (each cell represents a single scenario). For each cell of data, we had information on the simulation parameters and the results of several regression analyses. These analyses included tests of both correctly and incorrectly specified models. In particular, we regressed the group level outcome on climate level (B_1) and climate strength (B_2), and then on climate level (B_1), climate strength (B_2), and the interaction of climate level and climate strength (B_3) in a second regression. In each scenario, the regressions were repeated for each of the seven dispersion measures.

Analyses

Once the datasets were created and the measurement model for each measure of dispersion tested, we compared the average frequency with which significance was detected for various groupings of the simulation parameters. We also conducted a series of regression analyses to reveal the relative performance of the dispersion measures. We used logistic regression to predict whether the beta-coefficient associated with the dispersion measure correctly identified the effect as statistically significant. We ran separate regression analyses for each beta-coefficient (B_1, B_2, and B_3), and for when an effect was present or not. Thus, in all, we ran six logistic regressions.

As each simulation provided tests of the seven dispersion measures, we stacked the data. In total, we examined 6,988,800 outcomes (9,984 scenarios x 100 cells per scenario x 7 outcomes per cell). However, because we ran separate analyses for when an effect was present or when an effect was not present, the regressions had different sample sizes. When modeling the correct identification of significance for B_1, B_2, and B_3 when no effect was present (i.e., not significant), the sample sizes were 1,747,200, 2,150,400, and 1,747,200, respectively. When modeling the correct identification of significance when an effect was present (i.e., was significant), sample sizes for B_1, B_2, and B_3 were 5,241,600, 4,838,400, and 5,241,600, respectively. Note that the sample sizes are different when modeling B_2 because, as mentioned earlier, we do not have a perfectly factorial design. Again, this occurred because we eliminated simulations in which a climate strength effect was present but variability was zero.

As independent variables in our logistic regressions, we included both observable characteristics (i.e., characteristics of the dataset that a researcher would be able to measure) and the simulation parameters. For each simulated dataset, we computed ICC(1) and ICC(2). We included the number of employees per group (E), the number of groups (N_g), the base amount of individual variance explained by the group construct (δ_B), and the variability of variance for the group-to-individual relationship (δ_V). We also included the magnitude of effect sizes for climate level (β_1), climate strength (β_2), and the climate level by climate strength interaction (β_3). We chose to include all of these variables to control for variance attributable to the characteristics of the dataset being analyzed. Finally, the regression analyses included dummy variables representing each measure of dispersion[3]. By examining the significance of the dummy variables, we could assess when a particular measure was more (or less) likely to correctly identify significance.

Results

Correlations and eta coefficients between the dispersion indices and scatter plots of these relationships are included in Figure 3. Given that the correlations and etas are calculated from measures of dispersion for each group, multiple lines of data were obtained from each single simulation replication. Because of limitations on computing power, we used measures from 10 replications per simulation instead of the full 100. This resulted in a total sample size of 7,987,200 groups.

As shown in Figure 3, (absolute) correlations between the measures ranged from 0.796 to 1.000 (all significant at $p < .001$), with an average correlation of .912, thus showing that they are all very similar to each other. The coefficient of variation (V) had the lowest correlations with an average correlation of .811, suggesting that it differed most from the other indices. Excluding the coefficient of variation from the average correlation calculation increased the average association between the remaining dispersion measures to 0.943. Consistent with the findings of prior research, the correlations between the indices indicate that up to 36% of their variance (or up to 19% excluding the correlations with coefficient of variation – V) is not shared. For example, while the original and modified interrater agreement indices (r_{wg} and r^*_{wg}, respectively) were highly consistent, r^*_{wg} demonstrated less convergence with the normalized coefficient of variation (V”). The scatter plots included in the upper diagonal of Figure 3 clearly illustrate these differences in the relationships between the dispersion indices.

We also report eta values for each relationship in Figure 3. While correlations are based on an assumed linear relationship, eta does not make this assumption, and thus represents the strength of the relationship from the best-fitting smooth curve (Nunnally & Bernstein, 1994). As shown in the figure, r_{wg}, and $r*_{wg}$ are perfectly linearly related, which is indicated by their scatterplot and an eta value of 1.000. In contrast, r_{wg} and SD are perfectly nonlinearly related. Notably, the eta values in Figure 3 are higher than their respective correlations with an average eta value of 0.932 (or 0.976 excluding relationships with the coefficient to variation), thus demonstrating the similarities between the dispersion indices. However, such similarities do not necessarily imply that the measures will perform the same when used as independent variables in further analyses. For that reason, we compared the relative performance of the dispersion indices – specifically, their proclivity toward Type I and Type II errors.

Table 3 reports the average frequencies with which the regression coefficients for each dispersion measure were statistically significant. When no climate strength effect is present, we should not detect significance. Setting alpha at 0.05, we would expect false positives 5% of the time. As shown by the results, this is supported for the detection of a strength effect for all indices. However, the results demonstrate that the likelihood of detecting significance was greater (and larger than 0.05 at $p < .0001$) when a strength effect is present. Although the average probabilities across all dispersion indices were only 6% for a small strength effect, 10% for a medium strength effect, and 19% for a large strength effect, all of the indices detected significance more often than when a strength effect was not present. The coefficient of variation (V), however, was shown to be less likely to detect a true underlying relationship when considering strength effects.

We also examined the probability of detecting significance on the climate level effect. As shown in Table 3, a Type I error is likely to occur when no level effect is present, the probability of which is considerably higher than that for detecting a strength effect when no such effect is present. Because the frequencies for detecting significance on the level effect were dramatically higher than those for the strength effect, we suspected that the results may have been influenced by the misspecification of the model (i.e., when an interaction between climate level and climate strength was present but not modeled). To test this, we examined cases in which there was no interaction effect present. Specifically, we modeled the base case of no effects (i.e., the null hypothesis) in which no level, strength or interaction effects were present as well as a case in which large level and strength effects are present yet there is no interaction effect. As shown in table 3, when there are no effects, all of the measures detected significance as often as expected. Additional analyses showed that the confidence interval for the level and strength effect tests for all of the indices included 0.05. For the correctly specified large manipulation, the results show acceptable power (c.f. Cohen, 1992) for detecting level effects with the exception of the coefficient of variation, which was significantly lower than that for all other measures (and slightly lower than the 0.80 level recommended by Cohen, 1992). The likelihood of detecting strength effects was approximately 20% across all indices. Thus, the likelihood of detecting strength effects is notably lower than the likelihood of detecting level effects.

Table 4 shows the frequencies of statistical significance for the dispersion measures when modeling main effects and interactions. As shown by the results, the likelihood of detecting significance when no interaction effect was present was close to the expected 5%. However, the results also show that the likelihood of detecting significance only changed modestly when an effect should have been discovered. Frequencies for statistical significance were only marginally above 0.05 when the interaction effect was small. That is, while the probability of detecting significance was statistically significantly greater than 5% ($p < .0001$), the likelihood of detecting significance across the seven dispersion measures was only 6%. The likelihood of detecting significance was greater (and larger than 0.05 at $p < .0001$) when the interaction effect was medium or large although the average probabilities across all dispersion measures was only 9% and 15%, respectively. The results show that the frequency of statistical significance was significantly lower ($p < .001$) for the alternate interrater agreement index (a_{wg}), coefficient of variation (V) and normalized coefficient of variation (V”) when the interaction effect was large. However, the average probability across the other four dispersion indices was only 16%.

We also modeled the base case of no effects (i.e., the null hypothesis) in which no level, strength or interaction effects were present as well as a case in which all three effects were large (i.e., strong manipulation). As shown in table 4 for the null hypothesis case, all of the measures detected significance as often as expected. For the strong manipulation, the results show that all measures were more useful for detecting significance, with the frequencies of statistical significance being relatively higher for level effects. However, while all of the indices can detect level, strength and interaction effects, the power for detecting such effects is well below the 0.80-level suggested by Cohen (1992).

Tables 5 and 6 summarize the results of the logistic regression analyses. Table 5 reports the performance of each index (relative to the other indices) when no level, strength or interaction effects are present. As shown in the table, the alternative interrater agreement index (a_{wg}) performs relatively better than the other dispersion indices when no level effect is present. Specifically, a_{wg} is significantly less likely than all other measures to detect a level effect when no level effect is present. Also shown in the table, the coefficient of variation (V) performs significantly worse than all other dispersion measures when no level effect or interaction effect is present. There were no statistically significant differences between the measures when there is no strength effect.

Table 6 reports the relative performance of the indices when level, strength or interaction effects are present. As shown in the top section of the table, the revised interrater agreement index (r^*_{wg}) performs worse than the other measures when a level effect is present, while the coefficient of variation (V) performs better than the other measures under these circumstances. In other words, r^*_{wg} is significantly less likely to detect a level effect when such an effect is present, while V is significantly more likely to detect a level effect when one is present. Despite the performance of coefficient of variation (V) under such conditions, its relatively poor performance as compared to the other dispersion indices in the presence of an interaction effect is shown in the bottom section of the table. The results also reveal that the alternative interrater agreement index (a_{wg}) and normalized coefficient of variation (V”) underperform the other measures when a strength or interaction effect is present. In comparison, the average deviation index (AD), standard deviation (SD), and the interrater agreement index (r_{wg}) are all significantly more likely to detect an interaction effect when one is present. However, SD outperforms both AD and r_{wg} when there is a strength effect or an interaction effect.

Discussion

Dispersion, or variability in the individual scores of work unit members, has been conceptualized as a unit-level construct and examined as a predictor of unit-level outcomes (Chan, 1998; Kozlowski & Klein, 2000). Although researchers have utilized a variety of indices, for which methodological advantages and limitations have been identified (Bedeian & Mossholder, 2000; Brown & Hauenstein, 2005; Burke & Dunlap, 2002; Burke et al., 1999;

Kozlowski & Hattrup, 1992), little research has considered how the operationalization of dispersion may impact the inferences that may be made about the relationship between within-group variance and group-level outcome variables. The goals of this study were to explore the validity of various dispersion indices for detecting meaningful relationships between group member perceptions and outcome variables. Within the context of psychological climates, these goals were accomplished by examining the results from simulated data under conditions with variable sample characteristics, relationships between a group-level construct and individual-level measures of that construct, and relationships between the group-level construct and a group-level outcome.

Consistent with the findings of prior research (see Bedeian & Mossholder, 1998), the simulations revealed the weaknesses of coefficient of variation as a measure of dispersion. Specifically, the results highlighted a lower likelihood of detecting a true relationship between variability in group member perceptions and group-level outcomes when using coefficient of variation as a measure of dispersion. The results also showed that the coefficient of variation was less likely to detect interaction effects. Surprisingly, however, coefficient of variation outperformed other measures when detecting level effects. In particular, coefficient of variation was more likely than the other indices to detect significance when a level effect was present.

Interestingly, the normalized coefficient of variation did not show a marked improvement in performance over the coefficient of variation. Although normalizing the coefficient of variation adjusts for differences in group or sample sizes, the results of our study demonstrated that the measure has greater power than only the alternate index when modeling strength and interaction effects. Further, the normalized coefficient of variation performed significantly worse than all other measures (except for the alternate interrater agreement index) in the presence of medium to large interaction effects. Therefore, the usefulness of the normalized coefficient of variation for detecting meaningful relationships between variability in group member perceptions and group-level outcomes may be limited.

Beyond the unnormed and normed coefficients of variation, the results highlighted some differences between the other dispersion indices. For example, although the alternate interrater agreement index has several advantages over other measures given its independence from sample sizes, scales, or the location of observed means, the results demonstrated its relatively low performance in comparison to other dispersion measures when level, strength or interaction effects are present. Similarly, the revised index of interrater agreement only significantly outperformed the alternate index and normalized coefficient of variation when strength effects were modelled. Our findings suggest that standard and average deviation indices may perform better than interrater agreement indices when modeling level effects, and better than the coefficients of variation and the alternate interrater agreement index when modeling interaction effects. However, standard deviation has a higher likelihood of detecting significance when strength or interaction effects are present.

Despite these findings, the results also uncovered an interesting similarity between the dispersion indices – little statistical power for detecting strength and interaction effects. Specifically, the simulations showed that true relationships will be detected less than 30% of the time. This may help to explain the findings of prior research in which main effects of climate strength on group-level outcomes were not found (see Colquitt et al., 2002; Lindell & Brandt, 2000). Further, the frequencies of statistical significance in the presence of large interaction effects may provide insight into the mixed empirical results to date regarding climate strength as a moderator (see Gonzalez-Roma et al., 2002; Lindell & Brandt, 2000). More importantly, given that the likelihood of detecting strength and interaction effects was substantially lower than the 0.80-level suggested by Cohen (1992), our results also suggest that future research is needed to devise measures that more appropriately model dispersion.

Given the lack of statistical power of current dispersion measures, our results may also suggest that researchers exploring the effects of within-group variability may want to consider setting higher alpha levels (e.g., alpha = .10). Although establishing a higher alpha level is contrary to the conventional .05 standard, it might be helpful for addressing some of the general difficulties of detecting strength effects or interactions in moderated multiple regression. At the same time, however, a higher alpha level would by definition increase the likelihood of erroneously concluding the presence of significance. Thus, researchers should seriously consider the theoretical rationale for a strength effect as well as the potential implications of Type I and Type II errors in their specific contexts. Although a higher alpha level may not be appropriate for investigating strength effects in exploratory analyses, methodological issues associated with extant dispersion measures may merit a deviation from the current standard in hypothesis testing until better-performing measures are developed.

Our study has some limitations that may limit the generalizability of our findings. Because simulations are driven by assumptions, there is the possibility that the conditions modeled in our simulation do not accurately represent reality. Additional research is needed to explore the relationship between dispersion and group-level outcomes under alternate conditions, including deviations from normality and non-Likert response scales. Although we found consistent effects across various scenarios, we realize that the relationship between group-level constructs and individual-level measures of those constructs may also differ from that simulated. Thus, our understanding of the predictive effects of climate strength may be advanced through theorizing or typologies that articulate the ways in which variability influences outcome variables (similar to Chan, 1998). Without a better understanding of the “true” way in which dispersion (as a group-level construct) affects outcomes, it is difficult to determine how well our results mimic real situations.

Given the sample sizes of the datasets included in our simulation, the differences uncovered between the dispersion measures may be statistically but not practically significant. As shown by the results, there were high intercorrelations between the measures as well as similar patterns in their ability to detect strength and interaction effects. Thus, while researchers may have theoretical and interpretative reasons for choosing between dispersion measures, the results of this study suggests that they may yield only slight differences in the inferences made about the relationship between dispersion and group-level outcome variables. However, the results may also suggest that the portion of variance that is shared by the indices included here is not completely random, such that specific indices are capturing different relationships between variability in group member perceptions and outcome variables. Therefore, research is needed to develop indices of dispersion that may better estimate such relationships.

Despite these limitations, our tests do provide a useful means for comparing the relative performance of the dispersion indices. Our results highlight the relationships between various indices of dispersion and reveal some differences in the validity of the measures for detecting meaningful relationships between group member perceptions and group outcomes. By examining the qualities and efficacy of these indices across a variety of scenarios, our tests allowed us to identify the conditions under which specific dispersion measures might be most effective. The findings of our study suggest that when modeling level effects, and if there is sufficient theoretical evidence to support the absence of strength or interaction effects, the coefficient of variation is the best predictor. However, if there is evidence to suggest that there are strength or interaction effects, our results indicate that researchers may be better-served by using standard deviation as a dispersion measure. Although the average deviation index may be nearly as useful as standard deviation for detecting the interactive effects of level and strength on group-level outcomes, standard deviation has a higher likelihood of detecting strength effects. The ease with which the standard deviation index can be calculated further supports its usefulness as a dispersion index.

References

Allison, P. D. (1978). Measures of inequality. American Sociological Review, 43, 865-880.

Bedeian, A. G., & Mossholder, K. W. (2000). On the use of the coefficient of variation as a measure of diversity. Organizational Research Methods, 3, 285-297.

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

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

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

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

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

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

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

Cohen, A., Dovey, E., & Eick, U. (2001). Statistical properties of the $r_{wg(j)}$ index of agreement. Psychological Methods, 6, 297-310.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Colquitt, J. A., Noe, R. A., & Jackson, C. L. (2002). Justice in teams: Antecedents and Consequences of Procedural Justice Climate. Personnel Psychology, 58, 83-109.

Colquitt, J. A., Zapata-Phelan, C. & Roberson, Q. (2005). Justice in teams: A review of fairness effects in collective contexts. In J. J. Martocchio (Ed.), Literature Review and Agenda for Future Research. Research in Personnel and Human Resource Management (Vol. 24, pp. 53-94) Oxford, UK: Elsevier.

Conway, L. G., III, & Schaller, M. (1998). Methods for the measurement of consensual beliefs within groups. Group Dynamics: Theory, Research and Practice, 2, 241-252.

Gonzalez-Roma, V., Peiro, J. M., & Tordera, N. (2002). An examination of the antecedents and moderator influences of climate strength. Journal of Applied Psychology, 87, 465-473.

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). r_{wg}: an assessment of within-group agreement. Journal of Applied Psychology, 78, 306-309.

Klein, K. J., & Kozlowski, S. W. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. Organizational Research Methods, 3, 211-236.

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

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

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{wg}, $r*_{wg}$, $r_{wg(j)}$, and $r*_{wg(j)}$. Organizational Research Methods, 8, 128-138.

Lindell, M. K., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. Journal of Applied Psychology, 85, 331-348.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. Applied Psychological Measurement, 23, 127-135.

Martin, J. D., & Gray, L. N. (1971). Measurement of relative variation: Sociological examples. American Sociological Review, 36, 496-502.

Microsoft Corporation (1998). Microsoft Visual Basic 6.0 [Computer software].

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. Psychological Review, 80, 252-283.

Morgeson, F. P., & Hoffmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. Academy of Management Review, 8, 547-558.

Nunnally, J., & Bernstein, I. (1994). Psychometric Theory (3rd edition). New York: McGraw Hill.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. Journal of Applied Psychology, 74, 368-370.

Schneider, B., & Reichers, A. E. (1983). On the etiology of climates. Personnel Psychology, 36, 19-39.

Schneider, B., Salvaggio, A., & Subirats, M. (2002). Climate strength: A new direction for climate research. Journal of Applied Psychology, 87, 220-229.

Schriesheim, C. A., Cogliser, C. C., & Neider, L. L. (1995). Is it "trustworthy"? A multiple levels-of-analysis reexamination of an Ohio State leadership study, with implications for future research. Leadership Quarterly, 6, 111-145.

Smithson, M. (1982). On relative dispersion: A new solution for some old problems. Quality and Quantity, 16, 261-271.

Sturman, M. C. (2004). DataSim [Computer software]. Retrieved from http://www.people.cornell.edu/pages/mcs5/Pages/DataSimPage.htm

Biographical Information

Quinetta M. Roberson is an associate professor of human resource studies at Cornell University. Her research includes multilevel investigations of organizational justice – particularly team justice climates and other collective work contexts. Her work also examines climates for diversity and inclusion in organizations, and the link between diversity and bottom-line outcomes.

Michael C. Sturman (http://www.people.cornell.edu/pages/mcs5) is an Associate Professor of Cornell University's School of Hotel Administration. He received his Ph.D. in 1997 from Cornell University's School of Industrial and Labor Relations. His research examines human resource decision-making, the prediction of individual job performance over time, and the consequences of compensation decisions.

Tony L. Simons is an Associate Professor of Cornell University's School of Hotel Administration. His work focuses on leadership, teamwork, employee climate, and negotiations with emphasis on the concept of behavioral integrity, or the perceived pattern of alignment between managers' words and deeds.

Endnotes

[1] The calculation of the interrater agreement index differs based on whether one item per group member or multiple items per group member are used (James et al., 1984). Accordingly, we refer to r_{wg} when discussing issues relevant to both indices and to $r_{wg(I)}$ or $r_{wg(J)}$ when discussing issues pertinent to a specific measure. We did not, however, include number of items as a simulation parameter.

[2] We searched various reference sources for research that has examined relationships at the group or team level of analysis. Specifically, we looked for studies in which measures were taken at the individual level and within-group agreement indices were used to justify aggregation of those responses to represent scores at the group level. Reference sources for this review included: Academy of Management Journal, Administrative Science Quarterly, Journal of Applied Psychology, Journal of Management, Organizational Behavior and Human Decision Processes and Personnel Psychology. Overall, 23 studies were included in our analysis. A list of these studies and their sample characteristics may be obtained from the first author.

[3] Given that the focus of this study is on the relative performance of various dispersion measures, we do not report the regression results associated with the non-dummy independent variables. However, those analyses may be obtained from the second author.

Table 1

Formulae for Dispersion Measures

Standard Deviation	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$
Average Deviation Index	$AD_{M(J)} = \frac{\sum_{k=1}^{N}\left\lvert x_{JK} - \bar{x}_J \right\rvert}{N}$ where N is the number of judges or observations for an item J, and x_{JK} is the Kth judge's rating on item J
Interrater Agreement Index	$r_{wg(J)} = \frac{J\left[1-\left(\bar{s}_x^2 / s_{EU}^2\right)\right]}{J\left[1-\left(\bar{s}_x^2 / s_{EU}^2\right)\right]+\left(\bar{s}_x^2 / s_{EU}^2\right)}$ where J is the number of items in the scale, $\bar{s}_x^2$ is obtained average variance on the J items, and s_{EU}^2 is the variance of the uniform distribution
Revised Index of Interrater Agreement	$r^*_{wg(J)} = 1 - \frac{\bar{s}_x^2}{s_{EU}^2}$ where $\bar{s}_x^2$ is the obtained average variance of the items in the scale and s_{EU}^2 is the variance of the uniform distribution

Table 1

Formulae for Dispersion Measures

Alternative Interrater Agreement Index	$a_{wg(J)} = \frac{\sum\left(1 - \frac{2 * s_x^2}{\left[(H+L)M - \left(M^2\right) - \left(H * L\right)\right] * \left[K/(K-1)\right]}\right)}{J}$ where H is the maximum possible values of the scale, L is the minimum possible value of the scale, M is the observed mean rating, K is the number of raters, and J is the number of items in the scale
Coefficient of Variation	$V = \frac{s}{\overline{x}}$
Normalized Coefficient of Variation	$V'' = \left(\frac{\left(\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right)}{\left(\sum_{i=1}^{N} Q_i^2 - N\overline{X}^2\right)}\right)^{1/2}$ where Q_i equals the ith data value in a set of specified N values that produce the maximum variation in a response variable

Table 2

Simulation Parameters and Levels

Parameters	Levels in Simulation
Sample characteristics	
Employees per group (E)	3, 5, 10, 25
Number of groups (N_g)	40, 80, 120
Relationship between group-level construct and individual measures	
Base amount of individual variance explained by the group-level construct (δ_B)	0.00, 0.10, 0.30, 0.50
Variability of variance for group-to-individual relationship (δ_V)	0.00, 0.10, 0.30, 0.50
Relationship between group-level predictors and group-level outcome	
Effect of group-level construct, or climate level (β_1)	0.00, 0.10, 0.30, 0.50
Effect of total amount of individual variance explained by the group construct, or climate strength (β_2)	0.00, 0.10, 0.30, 0.50
Effect of climate level x climate strength interaction (β_3)	0.00, 0.10, 0.30, 0.50

Table 3

Frequency of Statistical Significance for Dispersion Measures Results without Interactions Modeled

	AD_M		SD		r_{wg}		$r*_{wg}$		a_{wg}		V		V’’	
	B_1	B_2	B_1	B_2	B_1	B_2	B_1	B_2	B_1	B_2	B_1	B_2	B_1	B_2
Level effect														
None*	.312		.312		.312		.312		.312		.262		.312	
Small	.334		.334		.334		.334		.334		.287		.334	
Medium	.580		.580		.580		.580		.580		.481		.580	
Large	.796		.796		.796		.796		.796		.717		.796	
Strength effect														
None*		.051		.051		.051		.051		.051		.053		.051
Small		.057		.057		.057		.057		.057		.053		.057
Medium		.106		.108		.104		.104		.099		.084		.102
Large		.198		.204		.197		.197		.182		.161		.188
No effects for level, strength, and interaction*	.050	.051	.050	.051	.051	.053	.051	.053	.049	.053	.050	.052	.050	.051
Large effects for level, strength, and interaction	.803	.237	.804	.244	.803	.238	.803	.238	.802	.221	.739	.165	.802	.225

Note. B_1 and B_2 are calculated effect sizes for level and strength, respectively. Numbers in the table represent the frequency with which significance was detected. For rows marked with *, significance indicates a Type I error with an alpha level of .05. In all other rows, a lack of significance indicates a Type II error, such that values represent the probability of correcting detecting a significant relationship (i.e., power) and should ideally be large.

Table 4
Frequency of Statistical Significance for Dispersion Measures Results with Interactions Modeled

	AD			SD			r_{wg}			r^*_{wg}			a_{wg}			V			V’’		
	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
Level effect																					
None*	.169			.169			.082			.077			.074			.174			.149		
Small	.159			.159			.092			.083			.075			.164			.140		
Medium	.173			.171			.133			.099			.110			.186			.153		
Large	.243			.238			.216			.125			.174			.301			.202		
Strength effect																					
None*		.051			.051			.050			.050			.050			.051			.049	
Small		.059			.059			.059			.059			.058			.060			.058	
Medium		.110			.114			.110			.110			.103			.171			.107	
Large		.200			.208			.201			.201			.184			.202			.190	
Interaction																					
None*			.051			.053			.051			.051			.055			.060			.054
Small			.058			.059			.059			.059			.057			.068			.057
Medium			.094			.096			.093			.093			.082			.093			.084
Large			.162			.167			.163			.163			.139			.135			.141
No effects*	.046	.040	.045	.052	.043	.047	.047	.042	.046	.044	.042	.047	.049	.044	.048	.049	.044	.044	.049	.046	.048
Large effects for level, strength, and interaction	.354	.242	.225	.349	.256	.233	.379	.248	.226	.252	.248	.226	.286	.230	.195	.383	.240	.233	.316	.242	.195

Notes. B_1, B_2, and B_3 are calculated effect sizes for level, strength, and their interaction, respectively. Numbers in the table represent the frequency with which significance was detected**.** For rows marked with *, significance indicates a Type I error with an alpha level of .05. In all other rows, a lack of significance indicates a Type II error, such that values represent the probability of correcting detecting a significant relationship (i.e., power) and should ideally be large.

Table 5

Relative Performance of Dispersion Measures when No Effect is Present

	AD	SD	r_{wg}	r^*_{wg}	a_{wg}	V	V’’
Level Effect (B_1)							
1. AD	n/a		-	-	-	+	-
2. SD		n/a	-	-	-	+	-
3. r_{wg}	+	+	n/a	-	-	+	+
4. r^*_{wg}	+	+	+	n/a	-	+	+
5. a_{wg}	+	+	+	+	n/a	+	+
6. V	-	-	-	-	-	n/a	-
7. V’’	+	+	-	-	-	+	n/a
Strength Effect (B_2)							
1. AD	n/a						
2. SD		n/a					
3. r_{wg}			n/a				
4. r^*_{wg}				n/a			
5. a_{wg}					n/a		
6. V						n/a	
7. V’’							n/a
Interaction Effect (B_3)							
1. AD	n/a					+	
2. SD		n/a				+	
3. r_{wg}			n/a			+	
4. r^*_{wg}				n/a		+	
5. a_{wg}					n/a	+	
6. V	-	-	-	-	-	n/a	-
7. V’’						+	n/a

Note. Dispersion indices listed vertically served as the baseline for comparison purposes. “+” indicates better (i.e., less likely to detect an effect) than the dispersion index listed horizontally, “-“ indicates worse (i.e., more likely to detect an effect) than the dispersion index listed horizontally, and an empty cell indicates no statistically significant difference between the two indices at $p < .05$.

Table 6

Relative Performance of Dispersion Measures when an Effect is Present

	AD	SD	r_{wg}	r^*_{wg}	a_{wg}	V	V’’
Level Effect (B_1)							
1. AD	n/a	+	+	+	+	-	+
2. SD	-	n/a	+	+	+	-	+
3. r_{wg}	-	-	n/a	+	+	-	-
4. r^*_{wg}	-	-	-	n/a	-	-	-
5. a_{wg}	-	-	-	+	n/a	-	-
6. V	+	+	+	+	+	n/a	+
7. V’’	-	-	+	+	+	-	n/a
Strength Effect (B_2)							
1. AD	n/a	-			+		+
2. SD	+	n/a	+	+	+	+	+
3. r_{wg}		-	n/a		+	-	+
4. r^*_{wg}		-		n/a	+	-	+
5. a_{wg}	-	-	-	-	n/a	-	-
6. V		-	+	+	+	n/a	+
7. V’’	-	-	-	-	+	-	n/a
Interaction Effect (B_3)							
1. AD	n/a	-			+	+	+
2. SD	+	n/a	+	+	+	+	+
3. r_{wg}		-	n/a		+	+	+
4. r^*_{wg}		-		n/a	+	+	+
5. a_{wg}	-	-	-	-	n/a	-	-
6. V	-	-	-	-	+	n/a	+
7. V’’	-	-	-	-	+	-	n/a

Note. Dispersion indices listed vertically served as the baseline for comparison purposes. “+” indicates better (i.e., more likely to detect an effect) than the dispersion index listed horizontally, “-“ indicates worse (i.e., less likely to detect an effect) than the dispersion index listed horizontally, and an empty cell indicates no statistically significant difference at $p < .05$.

Appendix A

Data Generation Algorithm

1. Determine scenario parameters (see Table 2)

   $E, N_g, \delta_B, \delta_V, \beta_1, \beta_2, \beta_3$

2. For each group, determine the total amount of variance in the individual variable to be explained by group membership

   $\delta_{Vi} = U(0, \delta_V)$

   Group Dispersion $(\delta_D) = 1 - (\delta_B + \delta_{Vi})$

   Note that the amount of dispersion $(\delta_D) = 1 - (\delta_B + \delta_{Vi})$, so that 0 means no variability, and 1 means maximum variability.

3. Transform (δ_D) into standardized units

   Based on a prior simulation, we observe that (δ_D) has a mean of 0.6714 and SD of 0.1967

   $\delta'_D = (\delta_D - 0.6714)/0.1967$.

4. Generate group-level predictor

   $G_j = N(0,1)$

5. Generate group-level outcome

   Outcome $= [\beta_1 * G_j + \beta_2 * \delta'_D + \beta_3 * (G_j * \delta'_D)] + [\mathrm{sqrt}(1 - (\beta_1^2 + \beta_2^2 + \beta_3^2)) * N(0,1)]$

6. Generate individual-level measure for each group member

   $X_{ij} = [\mathrm{sqrt}(1 - \delta'_D) * G_j] + [\mathrm{sqrt}(\delta'_D) * N(0,1)]$

   Repeat this step for all employees per group (E)

7. Change individual-level measure to 7-point scale

   $X'_{ij} = (X_{ij} * 1.4) + 4$, with minimum of 1 and maximum of 7

   Repeat this step for all employees per group (E)

8. Repeat steps 2-7 for all groups (N_g)

9. Calculate measure of group level, or mean X, of each group ($\bar{x}$)

10. Calculate dispersion measures for each group (AD, SD, r_{wg}, $r*_{wg}$, a_{wg}, V, V")

11. Run regression analysis

    $Y = B_0 + B_1 * \bar{x} + B_2 *$ dispersion measure $+ B_3 * \bar{x} *$ dispersion measure

    Repeat for all seven dispersion measures

12. Repeat steps 1-11 10 times

13. Repeat steps 1-12 for all desired combinations of scenario parameters