

Methods for the Measurement of Consensual Beliefs Within Groups

Lucian Gideon Conway III and Mark Schaller
University of British Columbia

The study of consensus in groups is fundamental to the understanding of group processes and the psychological experiences of individuals within groups. Measuring consensus in groups is tricky. This article reviews strengths and weaknesses of various methods for measuring the magnitude of consensus between persons on a single target belief. Considered are methods based on mean extremity, percentage agreement, dispersion, correlation, and spatial clustering. Specific advantages, limitations, and interpretational pitfalls are considered for each measure.

Despite a diversity of personalities, roles, and histories, people often share very similar beliefs. This fact of consensus has fundamental consequences on group processes and the psychological experience of individuals within groups. Indeed, the study of consensual common ground has played an important role in many areas of small group research (Hoyle & Crawford, 1994; see Baron, Kerr, & Miller, 1992, for a review). For instance, examination of consensus is central to inquiry into how groups come to have the particular members that they do (e.g., Gailbreath, Wagner, Moffett, & Hein, 1997). In addition, because small group norms are, by definition, consensually shared beliefs, tests of group norm conformity models (e.g., Prapavessis & Carron, 1997) often involve the measurement of consensus. The study of consensus in small groups is not merely an academic issue; it has implications for the health and happiness of individuals in real-life groups. It appears that, within peer group networks, there emerge consensual norms governing health-related behaviors such as binge eating (see Crandall, 1988); consequently, the impact of therapeutic interventions relevant to these behaviors may be enhanced by an understanding of the social

contagion processes operating within those groups.

Consensual beliefs are also fundamental to the dynamics of groups and populations of much larger scope. Stereotypes, for example, have the negative consequences that they do precisely because they are consensually shared by large populations (Haslam, 1997; Schaller & Conway, in press). Indeed, the process of stereotype threat that contributes to the under-performance of African Americans on certain intellectual tasks (Steele & Aronson, 1995) is based on the perception that specific stereotypical beliefs about African Americans are widely shared. More broadly, because culture is defined on the basis of consensual beliefs and behaviors (Latane, 1996; Triandis, 1996), an understanding of the origins of culturally shared beliefs requires some attention to consensus.

Our knowledge of the processes underlying consensual beliefs is still quite modest. This modesty has not gone unnoticed. Hardin and Higgins (1996, p. 29) observed that the study of shared reality has maintained "a ghostly presence . . . seemingly everywhere and nowhere at the same time." Other scientists have lamented the relative paucity of research devoted to understanding consensus in those beliefs that matter explicitly because they are consensual, such as group stereotypes (Schneider, 1996). There is evidence recently that psychologists are not merely recognizing these lacunae but that they are doing something about it. Recent edited books and special issues of several psychological journals have highlighted the study of socially shared cognition (e.g., Resnick, Levine, & Teasley, 1991)—a symptom that the study of shared beliefs is a growing concern.

Lucian Gideon Conway III and Mark Schaller, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada.

Preparation of this article was supported by funds from the Social Sciences and Humanities Research Council of Canada.

Correspondence concerning this article should be addressed to Lucian Gideon Conway III, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia, Canada V6T 1Z4. Electronic mail may be sent to lconway@interchange.ubc.ca.

The purpose of this article is to help facilitate that growth by critically reviewing various methods that may be used to measure consensus in groups. As in any line of research that has proceeded haphazardly, empirical investigations into the emergence of consensus are often methodologically idiosyncratic and ad hoc. Researchers may be tempted to borrow a procedure here, invent a measure there, and otherwise cobble together a plausible methodology. Our recent experiences in the area reveal that it is very easy to find or to create any number of intuitively reasonable measures of consensus. However, our experiences also reveal that when dealing with something as volatile as human cognition and interaction, any subtle methodological oversight can spell the difference between interpretable and uninterpretable empirical results.

Although this review is perhaps relevant primarily to researchers who address questions pertaining to the causes and consequences of shared beliefs, the sensitivity to the measurement of consensus may have practical value even to researchers who are not directly interested in consensus as a conceptual variable. The measurement of consensus is relevant to the interdependence of group members' responses—a methodological consideration in any empirical investigation involving interacting groups (e.g., Prapavessis & Carron, 1997; see also Burlingame, Kircher, & Honts, 1994). To the degree that within-group members' responses are consensual, relative to between-group members' responses, they are considered interdependent and statistically must be treated differently than independent responses. Given this, it is useful for group researchers to have a general understanding of the advantages, limitations, and pitfalls of the various methods for measuring consensus. In addition, assessments of interrater agreement, often used to validate the coding of research participants' responses, invariably involve the measurement of consensus (e.g., James, Demaree, & Wolf, 1993; Kozlowski & Hatrup, 1992). Thus, this article, although not explicitly concerned with statistical interdependence or interrater agreement issues, is potentially relevant to researchers who must deal with such issues.

In this article, we discuss the advantages and limitations of specific measurement approaches to consensus. In addition, we consider some

conceptual contexts within which the measurement of consensus might occur, and we discuss additional methodological issues that arise in these contexts.

Scope of This Review

We cannot claim that this review is exhaustive. We limit this review to methodologies relevant to the measure of actual consensus, but we do not address the study of perceived consensus. There is considerable research on processes underlying subjective perceptions of group cohesion (Festinger, 1950; Mullen & Copper, 1994) or group consensus (Krueger, 1998; Mullen & Hu, 1988). We do not attempt to review those measures. In addition, there are special methodological issues that need to be considered when drawing conclusions about differences between actual and perceived consensus (Dawes, 1989; Krueger, 1998). We do not cover these topics either.

In addition, when considering the conceptual context within which measures of consensus are assessed, we limit our discussion to contexts in which consensus emerges implicitly, in the absence of explicit directives to attain consensus. We do not consider contexts in which consensus is assumed but not measured or in which the achievement of consensus is an explicit goal of ad hoc groups (e.g., much research on group decision making).

Finally, we limit this review to the measurement of consensus on a single target variable (e.g., the degree to which two or more persons agree on one particular belief). We do not consider the more sophisticated methods that may be used when addressing questions concerning more holistic conceptualizations of consensus based on multiple attributes, such as political ideologies. Nor do we consider the more sophisticated methods that, within multiattribute contexts, allow one to determine the extent to which consensus reflects perceiver variables, target variables, and the interaction of perceiver and target variables (e.g., Albright, Kenny, & Malloy, 1988).

Methods for Measuring the Magnitude of Consensus

When researchers measure consensus within some focal set of individuals, they do so usually

because they are addressing one of two questions: "Is there consensus?" or "How much consensus is there?" These two questions involve two distinct meanings of the word *consensus*. The first question ("Is there consensus?") demands a yes or no answer and therefore requires that consensus be defined against some precise threshold of agreement. For example, if everyone agrees exactly, then consensus is judged to exist; if not, then no consensus exists. The second question ("How much consensus is there?") implies that consensus is measured as a quantity, for example, the degree of agreement among individuals. An answer to the first question necessarily demands an answer to the second question. For the purposes of this review, therefore, we define consensus in the second sense, referring to the amount of agreement among individuals within a focal population.

Various methods of measurement have been or might be used to assess consensus in various contexts. Depending on the conceptual context within which it is assessed, each has some advantages and some important limitations. Table 1 offers a summary of some of the basic attributes of the consensus measures reviewed here.

Mean Extremity Measures

In some contexts, it may be plausible to infer consensus from the mean extremity of attitudes expressed within a population. Such methods have been used occasionally in studies of group stereotypes (e.g., Brigham, 1971; see Gardner, 1994, for a critical discussion). The logic is as follows: Suppose one measures an attitude on a

bipolar scale, anchored by endpoints of extreme disagreement-agreement. The midpoint on the scale (e.g., a 5 on a 9-point scale) conceptually represents the lack of strong belief. Thus, the presence of a specific attitude can be inferred from the extent to which an individual's attitude deviates from that midpoint. It follows that greater extremity in mean attitude across a focal population must, in general, represent greater attitudinal consensus among those individuals.

Advantages. There are few advantages, if any, to this particular measure. Under some limited set of circumstances, extremity might serve as a rough proxy for consensus and so might afford interpretations about consensus under conditions in which more direct measures of consensus are unavailable. Under most circumstances, the limitations of this approach are more salient.

Limitations and pitfalls. One disadvantage to this method is that it can only serve as a rough indicator of consensus when beliefs are measured explicitly on a clear bipolar scale. This is very limiting. However, even under these conditions, the measure is conceptually hazardous: Mean extremity is not directly interpretable as an indicator of consensus, but rather as an indicator of a conceptually distinct construct that may be (but may not be) influenced by consensus. Although measures of mean extremity and consensus may be correlated under some naturally occurring conditions, this cannot be assumed to always be the case. The two variables are conceptually orthogonal and are empirically orthogonal under many circumstances as well (see Jackson, 1975). Thus, there is considerable opportunity for misinterpretation.

Imagine a population of 6 individuals in

Table 1
Some Attributes of Different Measures of Consensus

Measure	Useful for binary responses?	Useful for nonbinary responses?	Types-sizes of groups for which measure is useful	Ease of computation	Intuitive appeal of scale	Ease of comparison across studies
Mean extremity	No	Yes	Any group/ $N > 2$	Easy	Low	Variable
Percentage agreement	Yes	No	Any group/ $N > 2$	Easy	High	Easy
Variance-standard deviation	Yes	Yes	Any group/ $N > 1$	Easy	Low	Difficult
Variance ratios (e.g., r_{wg})	Yes	Yes	Any group/ $N > 1$	Moderate	High	Difficult
Pearson r	Yes	Yes	Nonexchangeable/ $N = 2$	Easy	High	Easy
Intraclass correlation	Yes	Yes	Exchangeable/ $N > 1$	Variable	High	Easy
Quasi-dyad r	Yes	Yes	Exchangeable/ $N > 2$	Difficult	High	Easy
Monte Carlo r	Yes	Yes	Exchangeable/ $N = 2$	Easy	High	Easy
Spatial clustering p	Yes	No	Spatially anchored/ $N > 2$	Difficult	Low	Difficult

which every individual indicates an attitude at exactly the scale midpoint—say a 5 on a 9-point scale. Now imagine a population in which 3 of 6 individuals report a 9 and the other 3 all report a 1. In both populations, the mean attitude reported is a 5, indicating lack of consensus. And yet, clearly, the amount and nature of agreement differs substantially in the two populations. The mean extremity measure fails to identify those differences.

In addition, when used as a proxy for consensus, mean extremity measures may also appear to reveal differences that do not really exist. Imagine a population of 6 individuals all of whom indicate a score of 7 on a 9-point scale. Now imagine a population of 6 individuals all of whom indicate a score of 8. In fact, both populations show perfect agreement amongst the individuals; but the mean extremity score erroneously implies greater consensus within the second population.

Percentage Agreement Measures

Perhaps the most intuitively appealing method for measuring consensus directly is calculation of the percentage of individuals within a population who endorse a particular belief. To the degree that the percentage approaches 100%, there is greater consensus. This measure of consensus is popular in both the stereotype literature (e.g., Gilbert, 1951; Karlins, Coffman, & Walters, 1969; Katz & Braly, 1933) and small group literature (e.g., Prapavessis & Carron, 1997; Gailbreath et al., 1997).

Advantages. Perhaps the primary advantages of the percent agreement method are its ease of calculation and ease of interpretation. Percentages are reported on an intuitively appealing 0–100 scale. To report a consensus of 90% is readily understood to indicate greater agreement than 70% consensus. Further, given the common scale of measurement, it is possible to compare consensus results from different investigations done in different locales or at different times. This feature has been central to the value of descriptive investigations into changes in ethnic stereotypes across time (e.g., Gilbert, 1951; Karlins et al., 1969).

Limitations and pitfalls. Although superficially simple to interpret, percentage measures may pose some subtle interpretational difficul-

ties. For instance, in many contexts, it is not always clear how to interpret values lower than 50%. Values closer to 0 may indicate less consensus, or they may indicate greater consensus on a belief logically opposite to that being assessed. Imagine a population in which 50% of individuals agreed with the statement, "Taxes are too high." Now imagine a population in which 10% of individuals agreed with the statement, "Taxes are too high." Does the second example illustrate a situation in which there is less consensus that taxes are too high, or does it illustrate a situation in which there is greater consensus that taxes are too low?

To answer that question demands attention not just to the measurement scale, but also to the manner in which responses were generated. Percentage measures are highly dependent upon the specific methods under which individuals' responses are generated. This method dependence is amply illustrated in several well-known descriptive studies of ethnic stereotypes (e.g., Karlins et al., 1969; Katz & Braly, 1933). In these studies, participants identified the set of five trait adjectives that best described a given ethnic group. On the basis of such responses, it is easy to calculate the percentage of individuals who endorse a particular trait description. These percentage values are highly dependent upon the number of choices that individuals were given. As Gardner (1994, p. 7) pointed out, "there is no way of knowing if an individual fails to select an adjective whether that individual feels that the adjective definitely isn't applicable or simply whether it isn't as appropriate as others." In general, percentage-based measures are relatively insensitive to the gradients of beliefs that people typically hold.

Moreover, percentage agreement estimates are conceptually appropriate only with beliefs that are truly dichotomous. These beliefs are rare. Most of the interesting constructs that psychological researchers wish to measure are not reliably captured with simple binary response scales.

Finally, percentage estimates are fairly insensitive when assessing consensus within focal populations of small size. For example, assessing percent agreement within dyads yields only three possible response values: 0%, 50%, and 100%. Thus, as the size of the focal group becomes smaller, the utility of the percentage estimates likewise decreases.

Dispersion Measures: Variance–Standard Deviation

In some circumstances, it may be desirable not to measure consensus, *per se*, but rather to measure lack of consensus or lack of agreement within a focal population. One appealing method is to calculate a descriptive statistic of dispersion, for example, variance or standard deviation.

Advantages. There are a number of advantages of using a dispersion statistic as a measure of consensus. First, in contrast to percentage measures, dispersion statistics are amenable to beliefs measured on nonbinary response scales. Second, these measures are relatively easy to calculate and are relatively easily understood by any audience familiar with basic statistical methods. Finally, these measures are usable with a wide range of group sizes—any population with two or more individual scores on the same measure lends itself to the calculation of a dispersion statistic.

Limitations and pitfalls. One modest disadvantage of dispersion statistics—particularly in comparison to percentage measures—is that they are not intuitively appealing to lay audiences. The metric is largely mysterious to individuals who have not had training in statistical methods (and sometimes even to those who have).

The interpretation of actual values is further limited by the lack of clear anchor points on the measurement scale. Although perfect agreement is anchored by a value of 0, the upper boundary on the measurement scale is defined by the values on the response scale; even if those values are known, it is difficult to immediately ascertain the maximum possible value of the dispersion statistic. (Quick, what is the maximum possible standard deviation for a population of 5 individuals responding on a 9-point response scale?) Thus, the actual values on the measurement scale are scale specific and are relatively meaningless except in an exact comparative context. This places limits on the comparison of consensus results from different empirical investigations. In addition to these limits, unless comparing across conditions in an exact comparative context, variance measures do not lend themselves easily to inferential statistical inquiries.

Statistical indices of dispersion (variance,

standard deviation) may also yield misleading interpretations if the wrong formula is used as the basis for calculation. As any student of statistics learns explicitly, different formulae are used to calculate an index of dispersion, depending on whether it is calculated from all the scores in a population of interest or whether it is estimated on the basis of a sample from that population. As any student learns implicitly, the sample formula is the one that is almost always used in psychological research (indeed, the sample formula is the default on most spreadsheet software). It is all too easy to assume that the sample formula is also appropriate for the present purpose. In fact, that is usually not the case. When calculating variance or standard deviation as a measure of consensus within a group, the purpose is purely descriptive, not inferential (the inferential steps come later, for instance, when drawing conclusions about different amounts of consensus under different conditions). Typically, the scores from the entire focal population are available (e.g., if I want to know the actual amount of belief consensus that has emerged in a particular 5-person group, I have available all 5 scores from that group); consequently, a population formula, not a sample formula, is appropriate.

This is not mere statistical fussiness. When used inappropriately, the use of a sample formula to calculate dispersion may have one of two distinct undesirable consequences on inferential interpretation. This is because the index of dispersion calculated through a sample formula is influenced not only by the actual degree of dispersion within the set of individual scores but also by the number of individual scores within that set. To illustrate, consider 4 individuals, half of whom score 2 and half of whom score 4 on some attitude measure. Treated as a population, this set of responses has a variance of 1.0; treated as a sample, it has a variance of 1.33. Now consider a set of 8 individuals, half of whom score 2 and half of whom score 4 on the same attitude measure. Conceptually, the amount of dispersion is identical to that within the original set of 4 individuals. Consistent with that sense, the formula for the variance of a population yields a value of 1.0—exactly as in the set of 4 individuals. However, if treated as a sample, the variance is 1.14—lower than the sample variance in the set of 4 individuals.

This distinction is immaterial in studies in

which focal groups are all of the same size; but in studies in which group sizes differ, the distinction is important. In some studies, group size may be a random variable of no conceptual importance (e.g., group sizes vary between 3 and 5 as a result of recruitment procedures). In these cases, the use of a sample formula to calculate variance within groups introduces error variance and decreases statistical power relevant to inferential judgment. Consequently, it may lead to the failure to detect effects that really do exist.

On the other hand, in some studies, group size is itself an experimental variable of conceptual interest (e.g., a study designed to test a hypothesis concerning the impact of group size on emerging consensus). In these cases, the use of a sample formula to calculate variance within groups introduces a subtle confound (the group size variable is confounded with method variance). Consequently, it may lead to the empirical appearance of effects that do not actually exist.

Dispersion Measures: Variance Ratio Indices

It is possible to compute a more sophisticated variance index that more neatly maps onto the concept of meaningful consensus. One such measure is the r_{wg} within-group agreement index (James, Demaree, & Wolf, 1984; James et al., 1993). The r_{wg} agreement index is calculated by dividing the actual variance within a group by an estimate of the amount of variance that would be expected by chance alone, and then by subtracting this value from 1 (for computational details and psychometric discussion, see James et al., 1984, 1993). The resulting score therefore estimates the degree to which observed similarity in responses is due to actual agreement between group members.

A conceptually similar variance ratio index can be calculated in the same manner, except for the use of an estimate of the maximum possible variance as the divisor (for computational details and discussion, see Green, 1998; Jackson, 1975).¹

Advantages. Variance ratio indices such as r_{wg} offer several advantages over simpler calculations of variance or standard deviation. First, the index values conform to a scale anchored by 0 and 1 (values of r_{wg} may some-

times fall outside of the 0–1 range, but under most circumstances, such values are unlikely; James et al., 1993; Kozlowski & Hattrup, 1992). Because higher values indicate greater agreement (unlike simple dispersion indices), these indices offer straightforward, intuitively appealing measures of consensus.

In addition, r_{wg} has a built-in control for chance agreement. Without controlling for chance agreement, it is not clear how much of the observed consensus emerged due to actual agreement between the participants and how much emerged due to chance. Although not important for comparison across conditions, this can be an important advantage for addressing the simpler question of whether consensus emerged in a given sample.

Limitations and pitfalls. These variance ratio indices are not without some limitations. They are—like measures of variance—affected by sample size. It can be difficult to interpret low agreement values under conditions where sample size is small (Kozlowski & Hattrup, 1992). This limitation is not trivial for researchers who are interested in very small groups or dyads.

In addition, when r_{wg} is calculated, the estimation of the expected variance is potentially complicated. There are many different methods for generating this estimation, and, as Kozlowski and Hattrup (1992, p. 166) note, “there is no commonly accepted theoretical model to justify selection of one possible expected distribution over another.” This not only poses a logistical problem for researchers in deciding how to calculate the expected variance, but it can cause difficulties in interpretation as well, especially when comparing across studies. Because the ultimate value of r_{wg} is dependent on the expected variance, it can only be realistically compared across studies in which the same criteria were used to generate expected variances. Thus, although the r_{wg} agreement index generally falls on a scale from 0 to 1, it is dangerous to blindly compare it across studies. Indeed, this danger is likely increased due to its intuitively appealing 0–1

¹ This index comparing actual variance to maximum possible variance was developed and has been used in contexts assessing within-group agreement across multiple attributes. Separate calculations of agreement on each individual attribute allow one to compute a mean agreement score that indicates the degree of crystallization across all relevant attributes (Jackson, 1975).

scale. Because the numbers seem anchored, it is tempting to assume cross-study comparability.

Last, variance ratio indices (like the variances from which they are computed) do not lend themselves easily to inferential statistical inquiries. The only way currently available to test null hypotheses pertaining to r_{wg} is through a Fortran program that estimates quantiles for the r_{wg} sampling distribution (Charnes & Schriesheim, 1995). Although useful, the program is limited to a maximum of 30 group members and uses confidence intervals to test for significance at the .01, .05, and .10 levels, and thus it does not yield exact p values.²

Correlation Coefficient Measures Among Nonexchangeable Individuals

There are several methods of calculating correlation coefficients that may offer reliable indicators of consensus within dyads, as long as there are multiple dyads within the data set of interest. The methods differ depending upon whether the dyad members are nonexchangeable or exchangeable (Griffin & Gonzalez, 1995).

The nonexchangeable case occurs when the dyad members differ systematically along some specific variable (e.g., the dyad members are chosen to differ on some classification variable, such as biological sex or some personality trait). Under such circumstances, one can systematically specify one dyad member's belief (e.g., that held by the man) as one variable and the other dyad member's belief (e.g., that held by the woman) as another variable, and one can calculate a Pearson's correlation coefficient indicating the relation between these variables across dyads.

Advantages. When multiple dyads are in evidence, these coefficients are easy to calculate. Moreover, the interpretation of consensus benefits from the fact that these coefficients are on a common, intuitively appealing scale.

One particular benefit of this measure when testing hypotheses about the causes of consensus is that it is not influenced by accidental agreement resulting from response biases common to all participants. For example, if there is any baseline tendency for participants to use only a restricted range on the response scale—and so indicate beliefs that appear close together—this will not artificially inflate the

value of the correlation coefficient. In comparison, the same baseline response bias can artificially increase consensus on dispersion measures. Therefore, to the extent that an obtained r is greater than 0, it can be interpreted as indicating some meaningful within-dyad similarity.

Furthermore, under some methodological circumstances, these coefficients offer a means of discerning whether the meaningful within-dyad similarity reflects some unique interpersonal interactions within the dyad or whether it reflects some commonality in purely individual-level cognitive processes. This distinction is often conceptually important. Under many conditions, consensus may emerge in the absence of any actual interaction between individuals, simply as a result of the fact that different individuals perceive a common informational input (Haslam, 1997). Therefore, when testing hypotheses about the influences of actual interpersonal interaction on the emergence of consensus, it is imperative to control for effects of common informational input. One way to do so is to experimentally manipulate the nature of the common informational input by creating systematic differences in the objective information encountered by different dyads. Zero-order correlations between dyad members' beliefs indicate the effects of both common informational input and unique interpersonal interactions within dyads. Partial correlations between dyad members' beliefs can also be calculated, controlling for the systemic differences in informational input. These partial correlations indicate consensus due solely to unique interactions within dyads. To the extent that these partial correlations exceed 0, they indicate that consensus has emerged, at least in part, because of those unique interactions.

Limitations and pitfalls. Despite the common scale underlying correlation coefficients, there are some potential difficulties in interpretation common to most uses of correlation coefficients. In theory, scores closer to 0 indicate less consensus. In fact, scores close to 0 may emerge as a result of restricted variability.

² Because r_{wg} has been used as an index of interrater reliability, several other more strictly psychometric limitations of the measure have previously been discussed within that specific context (e.g., James et al., 1993; Schmidt & Hunter, 1989).

Indeed, under conditions in which there exists perfect consensus both within dyads and across dyads (i.e., all individuals express exactly the same belief), the correlation coefficient will be 0. Similarly, values closer to 1 are interpreted as indicating greater consensus. And yet, it is possible to have an obtained value of 1.0 even under conditions in which there is no absolute agreement within any single dyad. This could occur under conditions in which there are systematic mean differences between the nonexchangeable dyad members (e.g., men generally report more positive attitudes than women toward some attitude object).

It is also difficult to interpret obtained values less than 0. If 0 indicates conceptually the total lack of consensus, then how does one interpret an obtained coefficient of $-.40$? Of course, under conditions in which consensus really is minimal, some negative correlations would be expected as a result of sampling error. It is also possible that negative correlations may reflect the operation of some psychologically meaningful process—a process that may have important implications toward understanding the emergence of consensus. The meaning of negative correlations almost certainly depends upon the idiosyncratic context of the given investigation.

Correlation Coefficient Measures Among Exchangeable Individuals

Many studies of dyads and other groups do not meet the nonexchangeable criterion. Instead, the members are exchangeable (Griffin & Gonzalez, 1995). This situation poses some interesting analytic problems that require unique solutions. Given that the members of the group are conceptually interchangeable, there is an inevitable arbitrariness to the exact order in which individuals' measured beliefs are entered into an array of data. And yet, the exact order may have a substantial influence on a correlation coefficient calculated from that array. Conceptually, the ideal solution to this problem would be to compute a correlation for every possible data entry combination and then to calculate the mean. In actuality, this solution is impractical. For any but the tiniest of samples, this method would involve the calculation of an impossibly large number of correlation coefficients. For example, in a study with only 4 exchangeable dyads, there are 16 different orders in which the

data might arbitrarily be entered. Fortunately, a conceptually similar estimate of consensus is provided by the calculation of the intraclass correlation (Donner & Koval, 1980; Fisher, 1925). Although it does not actually yield the same value as the mean correlation from all possible data-entry orderings, the value of the intraclass correlation itself is unaffected by the arbitrary orderings within exchangeable dyads or groups.

Calculation of the intraclass correlation is most straightforward when dealing with dyads. A relatively simple method of calculating a pairwise intraclass correlation coefficient for dyads is described in detail by Griffin and Gonzalez (1995).

It is also possible to calculate intraclass correlation coefficients for groups larger than $N = 2$. The computations for this coefficient are more complicated but yield an index that is conceptually and statistically identical in most (but not all) respects to the pairwise intraclass correlation coefficient (see Kenny & La Voie, 1985, for computational details and discussion).

Another method for calculating a correlation coefficient with larger (e.g., $N > 2$) groups is to convert the group into a series of artificial quasi-dyads, each of which is composed of (a) an individual member of the group and (b) the average of all the other members. For example, in a study on social contagion of binge eating (Crandall, 1988), group members' individual responses were correlated with the average of the responses within the relevant group (excluding the focal individual). A correlation coefficient can then be generated from these quasi-dyads in a manner similar to that described earlier for nonexchangeable dyads (for computational details and discussion, see Crandall, 1988).

Advantages. Once again, the computation of correlation coefficients offers some benefits by way of interpretation and communication. The values are anchored according to an intuitively appealing measurement scale and, because the index is scale-independent, can be compared across studies.

Intraclass correlations also offer some specific advantages. Given the problem to which they offer a solution, it is advantageous to have any measure available that produces the same value regardless of the initial order of data entry (which, in the exchangeable case, is always

arbitrary). When dyads are the focal population, the intraclass correlation is relatively easy to compute according to the methods described by Griffin and Gonzalez (1995).

Furthermore, as with correlations in the nonexchangeable case, intraclass correlation estimates of consensus are not artificially inflated by incidental sources of within-group similarity (e.g., scale usage biases). Similarly, the same methodological procedures may be used to disentangle the causal effects of common informational input and interpersonal interaction on the obtained index of consensus—in this case, by computing a partial intraclass correlation coefficient that controls for systematic differences in common informational input.

Limitations and pitfalls. These different means of arriving at a correlation coefficient suffer some of the same interpretational disadvantages of correlation coefficients described earlier. In addition, the specific measures have certain unique limitations as well.

For instance, the intraclass correlation coefficient is descriptively conservative: It generally yields a value that is lower than the mean of all possible data-entry combinations. To illustrate, consider a sample of 4 dyads in which the attitude responses within each dyad are as follows: (3, 2), (1, 3), (4, 5), (4, 3). Given these data, there are 16 different equally meaningful orderings in which the data might be entered for statistical analysis. Pearson correlations calculated on each of the 16 arrays range from .375 to .944, with a mean Pearson's $r = .608$. In comparison, regardless of data-entry order, the intraclass correlation coefficient = .356. Except in unusual or extreme cases, the computation of an intraclass correlation yields a value that may fail to adequately convey the full magnitude of agreement within the focal populations (this is especially so, because the values are represented on the familiar scale of the Pearson correlation coefficient).³

Another disadvantage of the intraclass correlation coefficient pertains specifically to inferential statistical contexts. When an obtained intraclass correlation is tested against a null hypothesis, the computation can be complicated (see Griffin & Gonzalez, 1995). A similar pragmatic disadvantage arises when using the quasi-dyad approach described by Crandall (1988). If the researcher's goal is to test the

obtained correlation against a null hypothesis of "no agreement whatsoever," the specific value of the null hypothesis is not exactly 0 and so requires some additional ad hoc statistical maneuvers to compute the appropriate value to associate with the null hypothesis (for a discussion, see Crandall, 1988).

Spatial Clustering Measures

The subjective perception of consensus within a population depends upon the breadth of the population examined. Within any large population of individuals, there may superficially appear to be little evidence of consensual patterns of belief. However, closer examination may reveal categorically distinct subpopulations, within each of which may be high degrees of consensus on subpopulation-specific beliefs. The emergence of multiple consensual subgroups within a bigger population is a defining element of coalition formation in working groups. It is also fundamental to the emergence of perceptually distinct "cultures" across social space (Latane, 1996).

In testing hypotheses about the emergence of consensual subgroups within a social geography, the consensus measures discussed earlier are largely useless. Rather, what is required is a measure that assesses coalescence—the degree to which beliefs become clustered in a predictable manner across social space. To the extent that such clustering emerges, an argument can be made that there is evidence of emergent multiple pockets of spatially anchored consensus.

This particular domain of inquiry is quite new

³ One possible solution to the underestimation problem associated with intraclass correlation coefficients uses Monte Carlo methods to compute an estimated Pearson's r for exchangeable dyads. Dyad members' responses can be entered into a statistical software spreadsheet, and the computer can be programmed to (a) generate a finite set of permutations of the data set, (b) calculate Pearson's r on the basis of each permutation, and (c) compute the mean Pearson's r from this sample. Our own preliminary investigations suggest that modest-sized samples of computer-generated permutations yield fairly reliable estimates of the mean r that is calculated from all possible permutations. The basic hardware and software available for most contemporary personal computers (e.g., an Intel Pentium computer chip and Microsoft Excel 7.0 spreadsheet software) make the calculation of this Monte Carlo r quite easy. (One example of Excel 7.0 programming code that computes a Monte Carlo r may be obtained from Lucian G. Conway III upon request.)

within experimental psychology, and so there are only a few extant examples of methods for assessing the emergence of spatial clustering of beliefs. One method for assessing clustering is described by Latane and L'Herrou (1996). According to this method, the beliefs of individuals can be compared in binary fashion (agree vs. disagree) with the beliefs of each individuals' neighbors (those individuals who are immediately proximal in social space). The total number of actual agreements can be compared to the distribution of agreements that would be expected simply as a result of chance (i.e., random spatial distribution of observed individual attitudes). The result is a probability value indicating the likelihood that the observed degree of clustering would have emerged simply as a result of chance—lower values of p indicate greater clustering (see Latane & L'Herrou, 1996, for details).

Advantages. The primary advantage of this measure is that it solves a difficult problem that is not solved by any of the other measures of consensus. Consequently, this method may detect the emergence of multiple psychologically and sociologically meaningful pockets of spatially anchored consensus under conditions in which other measures would simply indicate that overall consensus is low or nonexistent. If the experimental study of group processes is to shed conceptual light on the emergence of culturally shared beliefs under the dynamic conditions in which the boundaries of relevant subpopulations are themselves determined by the content of those beliefs, then methods of this sort are a necessary scientific tool.

Limitations and pitfalls. Nevertheless, there are some limitations to the clustering index. It is labor intensive to compute, because it requires the determination of the chance number of agreements between spatial neighbors and the associated distribution. The specific nature of this distribution differs depending upon (a) the number of individuals within the spatial population under study and (b) the specific social geometry of that population. Consequently, any given empirical investigation requires some sophisticated ad hoc statistical maneuvers to generate the clustering index. Given this context dependence, there is no single probability value that indicates perfect clustering. This makes it difficult to interpret the index as a straightforward indicator of the magnitude of spatial

clustering. It also limits the comparability of results obtained across different empirical investigations.⁴

Another limitation of clustering measures such as those used by Latane and L'Herrou (1996) is that they are useful only for dichotomous measures of individual belief. Finally, the value of clustering measures is limited to very specific questions about the coalescence of consensual subpopulations within a broader social geography. They are not particularly useful if the goal is simply to measure the magnitude of consensus that emerges within a single well-defined group.

Conclusions

Given the inherent messiness of transforming the dynamic processes of group interaction into quantitative values, there are plenty of difficulties in detecting signals of conceptual truth amid the unavoidable noise. Thus, perhaps more so than in many other fields of inquiry, the methods we use to measure emerging consensus exert important influences on the conclusions we are able to draw from the resulting data. As we have reviewed, specific methods of measuring consensus afford specific advantages and specific limitations and demand attention to specific interpretational issues. To the extent that researchers who study consensus ignore those issues, interpretation problems or inferential errors are likely to arise. Some of these inferential errors may be of the Type I variety, wherein we erroneously believe in the existence of some phenomenon that does not really exist. Eventually, of course, such errors of commission are likely to be righted through the self-correcting mechanisms of collective scientific inquiry. More perniciously, careless methodologies in the study of consensus are likely to

⁴ In the memory literature, there is a popular index of semantic clustering in free recall that is unaffected by the size of the recall list and that conforms to a user-friendly measurement scale: Perfect clustering is indicated by a score of 1 and random clustering by 0 (Roemer, Thompson, & Brown, 1971). Although adequate for assessing clustering in one-dimensional space, this index does not generalize to the considerably more difficult task of measuring clustering in two-dimensional space. It is plausible that some analogous metric might be developed for application to the present context; if so, it would be a valuable methodological contribution.

lead also to inferential errors of the Type II variety, in which we fail to detect phenomena that do exist. These errors of omission are not so easily corrected.

Just as our research methods impact the conclusions we draw, these methods influence the theories we produce (Frey, 1994). For psychological knowledge into the processes underlying the emergence of consensual beliefs to progress, it will be imperative for us to choose and use available methods that best match our specific research questions and to develop new methods that are even better.

References

- Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *55*, 387-395.
- Baron, R. S., Kerr, N. L., & Miller, N. (1992). *Group process, group decision, group action*. Pacific Grove, CA: Brooks/Cole.
- Brigham, J. C. (1971). Ethnic stereotypes. *Psychological Bulletin*, *76*, 15-33.
- Burlingame, G. M., Kircher, J. C., & Honts, C. R. (1994). Analysis of variance versus bootstrap procedures for analyzing dependent observations in small group research. *Small Group Research*, *25*, 486-501.
- Charnes, J. M., & Schriesheim, C. A. (1995). Estimation of quantiles for the sampling distribution of the r_{wg} within-group agreement index. *Educational and Psychological Measurement*, *55*, 588-598.
- Crandall, C. S. (1988). Social contagion of binge eating. *Journal of Personality and Social Psychology*, *55*, 588-598.
- Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*, 1-17.
- Donner, A., & Koval, J. J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics*, *36*, 19-25.
- Festinger, L. (1950). Informal social communication. *Psychological Review*, *57*, 271-282.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.
- Frey, L. R. (1994). The naturalistic paradigm: Studying small groups in the postmodern era. *Small Group Research*, *25*, 551-577.
- Gailbreath, R. D., Wagner, S. L., Moffett, R. G., III, & Hein, M. B. (1997). Homogeneity in behavioral preference among U.S. army leaders. *Group Dynamics: Theory, Research, and Practice*, *1*, 222-230.
- Gardner, R. C. (1994). Stereotypes as consensual beliefs. In M. P. Zanna & J. M. Olson (Eds.), *The psychology of prejudice: The Ontario symposium* (Vol. 7, pp. 1-31). Hillsdale, NJ: Erlbaum.
- Gilbert, G. M. (1951). Stereotype persistence and change among college students. *Journal of Abnormal and Social Psychology*, *46*, 245-254.
- Green, C. W. (1998). Normative influence on the acceptance of information technology: Measurement and effects. *Small Group Research*, *29*, 85-123.
- Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin*, *118*, 430-439.
- Hardin, C., & Higgins, E. T. (1996). Shared reality: How social verification makes the subjective objective. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 3). New York: Guilford.
- Haslam, S. A. (1997). Stereotyping and social influence: Foundations of stereotype consensus. In R. Spears, P. J. Oakes, N. Ellemers, & S. A. Haslam (Eds.), *The social psychology of stereotyping and group life* (pp. 119-143). Oxford: Blackwell.
- Hoyle, R. H., & Crawford, A. M. (1994). Use of individual-level data to investigate group phenomena: Issues and strategies. *Small Group Research*, *25*, 464-485.
- Jackson, J. (1975). Normative power and conflict potential. *Sociological Methods and Research*, *4*, 237-263.
- 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 interrater agreement. *Journal of Applied Psychology*, *78*, 306-309.
- Karlins, M., Coffman, T. L., & Walters, G. (1969). On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, *13*, 1-16.
- Katz, D., & Braly, K. (1933). Racial stereotypes in one hundred college students. *Journal of Abnormal and Social Psychology*, *28*, 280-290.
- Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology*, *48*, 339-348.
- 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.
- Krueger, J. (1998). On the perception of social consensus. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 163-240). San Diego: Academic Press.

Latane, B. (1996). Dynamic social impact: The creation of culture by communication. *Journal of Communication, 46*(4), 13-25.

Latane, B., & L'Herrou, T. (1996). Spatial clustering in the conformity game: Dynamic social impact in electronic groups. *Journal of Personality and Social Psychology, 70*, 1218-1230.

Mullen, B., & Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin, 115*, 210-227.

Mullen, B., & Hu, L. (1988). Social projection as a function of cognitive mechanisms: Two meta-analytic integrations. *British Journal of Social Psychology, 27*, 333-356.

Prapavessis, H., & Carron, A. V. (1997). Sacrifice, cohesion, and conformity to norms in sport teams. *Group Dynamics: Theory, Research, and Practice, 1*, 231-240.

Resnick, L. B., Levine, J. M., & Teasley, S. D. (1991). *Perspectives on socially shared condition*. Washington, DC: American Psychological Association.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation

of clustering in free recall. *Psychological Bulletin, 76*, 45-48.

Schaller, M., & Conway, L. G., III. (in press). From cognition to culture: The origins of stereotypes that really matter. In G. Moscovitz (Ed.), *Future directions in social cognition*. Mahwah, NJ: Erlbaum.

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, D. J. (1996). Modern stereotype research: Unfinished business. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Foundations of stereotypes and stereotyping* (pp. 419-453). New York: Guilford.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

Triandis, H. (1996). The psychological measurement of cultural syndromes. *American Psychologist, 51*, 407-415.

UNITED STATES POSTAL SERVICE
Statement of Ownership, Management, and Circulation (Required by 39 USC 2086)

1. Publication Title: *Group Dynamics: Theory, Research, and Practice* (ISSN) 1 0 8 9 - 7 6 6 9

2. Issue Date: October 1998

3. Filing Date: October 1998

4. Issue Frequency: Quarterly

5. Number of Issues Published Annually: 4

6. Annual Subscription Price: \$35/indiv., \$90/institution

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4):
750 First Street, NE - Washington, DC 20002-4242

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer):
750 First Street, NE - Washington, DC 20002-4242

9. Full Name and Complete Mailing Address of Publisher, Editor, and Managing Editor (Do not leave blank):
Educational Publishing Foundation/APA
750 First Street, NE
Washington, DC 20002-4242

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a trust or other legal entity, give its name and address.)
Susan Knapp, American Psychological Association
750 First Street, NE - Washington, DC 20002-4242

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box: None

12. Tax Status (For completion by nonprofit organizations authorized to mail at special rates) (Check one):
 Has Not Changed During Preceding 12 Months
 Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

13. Publication Title	14. Issue Date for Circulation Data Below	Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
Group Dynamics: Theory, Research, and Practice	September 1998		
15. Extent and Nature of Circulation			
a. Total Number of Copies (Net press run)		2,725	2,537
b. Paid and/or Requested Circulation			
(1) Sales Through Dealers and Carriers, Street Vendors, and Counter Sales (Not Mailed)		-----	-----
(2) Paid or Requested Mail Subscriptions (Include advance proof copies and exchange copies)		1,258	1,389
c. Total Paid and/or Requested Circulation (Sum of 15b(1) and 15b(2))		1,258	1,389
d. Free Distribution by Mail (Carriers, complimentary, and other free)		4	16
e. Free Distribution Outside the Mail (Carriers or other means)		-----	-----
f. Total Free Distribution (Sum of 15d and 15e)		4	16
g. Total Distribution (Sum of 15c and 15f)		1,262	1,405
h. Copies not Distributed			
(1) Office Use, Leftovers, Spoiled		1,463	1,132
(2) Returns from News Agents		-----	-----
i. Total (Sum of 15g, 15h(1), and 15h(2))		2,725	2,537
Percent Paid and/or Requested Circulation (15c/15i x 100)		99.7	99.0
16. Publication of Statement of Ownership <input type="checkbox"/> Publication required. Will be printed in the <u>December 1998</u> issue of this publication. <input type="checkbox"/> Publication not required.			
17. Signature and Title of Editor, Publisher, Business Manager, or Owner		Date	
Susan Knapp, Director		10/1/98	

Instructions to Publishers

- Complete and file one copy of this form with your postmaster annually on or before October 1. Keep a copy of the completed form for your records.
- In cases where the stockholder or security holder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting. Also include the names and addresses of individuals who are stockholders who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing corporation. In item 11, if none, check the box. Use item 11b if more space is required.
- Be sure to furnish all circulation information called for in item 15. Free circulation must be shown in items 15d, e, and f.
- If the publication had second-class authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published. It must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October.
- In item 16, indicate the date of the issue in which this Statement of Ownership, Management, and Circulation will be published.
- Item 17 must be signed.
- Failure to file or publish a statement of ownership may lead to suspension of second-class authorization.

PS Form 3526, September 1985 (Revised)