

Organizational Research Methods

<http://orm.sagepub.com>

On the Use of Likert-Type Scales in Multilevel Data: Influence on Aggregate Variables

Daniel J. Beal and Jeremy F. Dawson

Organizational Research Methods 2007; 10; 657 originally published online Jul 23, 2007;

DOI: 10.1177/1094428106295492

The online version of this article can be found at:

<http://orm.sagepub.com/cgi/content/abstract/10/4/657>

Published by:



<http://www.sagepublications.com>

On behalf of:



[The Research Methods Division of The Academy of Management](#)

Additional services and information for *Organizational Research Methods* can be found at:

Email Alerts: <http://orm.sagepub.com/cgi/alerts>

Subscriptions: <http://orm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations <http://orm.sagepub.com/cgi/content/refs/10/4/657>

On the Use of Likert-Type Scales in Multilevel Data

Influence on Aggregate Variables

Daniel J. Beal

Rice University

Jeremy F. Dawson

Aston University

In multilevel analyses, problems may arise when using Likert-type scales at the lowest level of analysis. Specifically, increases in variance should lead to greater censoring for the groups whose true scores fall at either end of the distribution. The current study used simulation methods to examine the influence of single-item Likert-type scale usage on ICC(1), ICC(2), and group-level correlations. Results revealed substantial underestimation of ICC(1) when using Likert-type scales with common response formats (e.g., 5 points). ICC(2) and group-level correlations were also underestimated, but to a lesser extent. Finally, the magnitude of underestimation was driven in large part to an interaction between Likert-type scale usage and the amounts of within- and between-group variance.

Keywords: *multilevel; aggregation; Likert-type scale; intraclass correlation; Monte Carlo*

In almost all studies of individuals nested within groups or organizations, researchers must contend with some form of aggregation of the individual-level responses. Methodological concerns over when such aggregation is appropriate and when it might not be appropriate have appeared with regularity during the past quarter century (e.g., Bliese, 2000; Chan, 1998; George & James, 1993; James, 1982; Jones & James, 1979). In addition, many multilevel researchers have been concerned about whether constructs are isomorphic at different levels of analysis and the appropriate means to test such notions (e.g., Chen, Bliese, & Mathieu, 2005).

Of course, proper conclusions for both of these important issues depend on the extent to which the data conform to the assumptions of the underlying statistical model. For example, if the data deviate from normality or exhibit heteroscedasticity (see Micceri, 1989, for a comment on this likelihood), then it is possible that appropriate decisions regarding aggregation or isomorphism may be threatened. Furthermore, when higher-level data are aggregations of lower-level variables, the implications of these assumptions become an increasingly complex issue because of the reliance of the higher-level distribution on the features of the lower-level data.

Authors' Note: The authors would like to thank Kelly de Chermont for her help with an earlier version of this article. Portions of this article were presented at the 2005 Academy of Management conference, Honolulu, Hawaii. Correspondence concerning this article should be addressed to Daniel J. Beal, Department of Psychology-MS 205, Rice University, 6100 Main St., Houston, TX 77005; e-mail: dbeal@rice.edu.

Relatively few papers, however, have considered how interpretations of higher-level variables may change because of problems that occur with lower-level data. Perhaps the primary reason for not considering these effects arises from the assumption that variance at each level is independent of other levels. Such an assumption suggests that problems occurring at one level may interfere with the ability to find conceptual or empirical relations between the levels, but such problems should not affect the actual interpretation of variables assessed at each level. There is, however, some research suggesting that these assumptions may not be tenable. For example, Bliese (1998) demonstrated that small samples of individuals within groups weakened the reliability of group-level scores, resulting in attenuated correlations with other group-level variables. Thus, a characteristic of the lower-level data can have a direct bearing on the amount of random error variance present in the aggregated higher-level variable. The current article describes another such effect, involving the use of Likert-type scales at one level of analysis and aggregations of these scales at another, higher level of analysis.

Use of Likert-Type Scales in Multilevel Organizational Data

Quite often, particularly in organizational research, data collected at the lowest level of analysis use some form of discrete ordinal or interval scale to represent variables that have a continuous underlying distribution (Kotz, Balakrishnan, & Johnson, 2000). Several areas of research have, in the past, considered the implications within a single level of analysis of using these response formats and the consequences that arise as a result (Komorita & Graham, 1965; Krieg, 1999; Landy & Farr, 1980; Russell & Bobko, 1992). Generally, these studies have shown that as the set of response options depart from the continuous underlying distribution (i.e., relatively fewer response options on the scale), statistical problems, such as unreliability and increased Type II error rates, occur. It also appears that there are diminishing returns for adding response options beyond a certain point, with studies often suggesting that seven response options as optimal (Cicchetti, Showalter, & Tyrer, 1985; Rasmussen, 1989). Despite these suggestions, many Likert-type scales still employ comparatively few response options, occasionally even as few as two (e.g., Gangestad & Snyder, 2000).

In the current study, we were concerned with the consequences of using such scales for higher-level constructs that are represented by aggregations of these lower-level variables across members of each group. Specifically, if the underlying distribution is represented using some form of Likert-type scale or the distribution is otherwise discretized, what effect does this have on our interpretation of aggregated versions of these scales? In the current article, we investigate these effects for single-item measures, as these are still quite common and allow us to limit the number of factors simultaneously examined. We return to the issue of using multi-item scales again in the discussion section.

There are, of course, multiple parameters of aggregated variables that could be affected by the use of Likert-type scales. For example, multilevel researchers may wish to consider the accuracy of the aggregations themselves (e.g., the mean score for the group), the variance of the aggregations, the reliability of the aggregations (i.e., ICC(2)), the proportion of total variance that occurs between the groups (i.e., ICC(1)), and, ultimately, correlations of these aggregations

with other group-level variables. We will consider each of these aspects of aggregate variables in turn. First, however, we will discuss potential reasons why using Likert-type scales at the lower level of analysis might have an influence on aggregate variables.

To begin, let us first discuss a simple two-level model of people nested within organizations. Most researchers have described the individuals within such organizations using the following model (e.g., Hofmann, Griffin, & Gavin, 2000; Hox, 2002; Raudenbush & Bryk, 2001):

$$x_{ij} = x_j + e_{ij}$$

This model assumes that each person's true score (x_{ij}) is a function of his or her particular organization's true score (x_j) on the variable in question plus some amount of individual-level error (e_{ij}). If x_{ij} reflects some sort of perceptual variable, such as climate for satisfaction for an employee at a bank branch, then each individual employee's true level of endorsement for satisfaction will be a function of the particular branch's level of satisfaction, plus some amount of individual experiential and perceptual variation. Typically, the amount of individual variation within a given branch reflects agreement for the satisfaction climate (LeBreton, James, & Lindell, 2005).

Given that some branches will have true scores that are relatively high or low on the continuum of climate for satisfaction, we would expect individuals, on average, also to exhibit relatively high or low scores on the continuum of satisfaction. This is not an unusual notion; people who work at less satisfying organizations have scores that reflect the less than satisfying elements of those organizations, some are more satisfied than the mean, some are much less satisfied. It is entirely possible, for example, for an organization that is very low on this dimension to have little to no overlap in its individual scores with an organization that is very high on this dimension.

Herein lies the potential problem when these scores are made discrete in some way. To continue with the example, let us say that satisfaction is assessed using an item such as, "I am satisfied with this organization," and that responses to this item range from 1 (*strongly disagree*) to 5 (*strongly agree*). If an organization's true score on this variable is, say, particularly low, then the observed distribution is likely to consist of many 1s, with relatively few higher responses. The underlying distribution of these responses, however, will not exhibit such a floor effect; rather, it will reflect the true variation of all those individuals whose scores fell into the 1 category. The result is that the true underlying distribution is unlikely to resemble the observed Likert-type scale distribution.

Indeed, these observed distributions have plagued researchers for many years and often are discussed as one type of truncated distribution called a censored distribution (Gupta, 1952; Pearson & Lee, 1908). Specifically, censored normal distributions are those in which the underlying distribution should be normal, but for some reason, some portion of the scores is not allowed to take on their true values. For example, if one were making behavioral observations of time spent socializing during work hours, one might assign the highest rating to any episode of socializing lasting an hour or more. In truth, the episodes that were included in the highest category are not identical, but the measurement censors any variance above an hour, thereby forcing them to obtain the same rating. In the current situation, some organizations (i.e., those whose true scores fall toward either end of the continuum of a given variable) will exhibit some amount of censoring, whereas other

organizations (i.e., those that fall toward the middle of the underlying distribution) will not. So there is no set amount of censoring that occurs for all groups of individuals; rather, censoring increases as the group-level true scores depart from the overall mean across groups.

This general effect is also likely to be exacerbated as the overall true variance of the responses increases. That is, as the total amount of variance begins to cover more of the response scale, a greater number of true scores will fall in the lowest and highest response categories. Note that this increased amount of censoring will occur when the between-group variance increases, when the within-group variance increases, or when both sources of variance increase. In the current study, we simulated data to model these situations, independently manipulating the amount of within- and between-group variance. This allowed us to systematically examine whether the censoring effect increased more with changes of within variance, between variance, or a combination of both.

Implications of Likert-Type Scale Usage on Aggregate Variables

It is our contention that the censoring effects described above have a systematic influence on some features of higher-level aggregated variables. First, the literature on censored distributions (Maddala, 1983) makes it clear that the observed means for organizations will be increasingly biased (i.e., less extreme) as their true means approach either end of the continuum. This effect occurs because lower (or higher) scores are not allowed to take on their true, very low (or very high) values. Similarly, when considering the expected value of the grand mean, we would expect Likert-type scale versions to be less extreme than true scores as well, with the exception of when it is in the middle of the scale, in which case it will be unbiased. Put differently, if the same variable was assessed both on a true continuous distribution (e.g., $M = 0$, $SD = 1$) and on a 5-point Likert-type scale (e.g., $M = 3$, $SD = 0.5$), we would expect to observe several things. First, groups with high or low true scores would have Likert-type scale means relatively closer to the mean of all groups when compared to the truly continuous scale. Second, if variance were to increase (e.g., the SD on the Likert-type scale increased to 1.0), the inward shifting of the means would increase. Third, because the grand mean is exactly at the scale midpoint, we would expect the long-term mean of all groups to be accurate for both the continuous scale and the Likert-type scale (i.e., because biased low and high Likert-type scale group means would cancel each other out); otherwise, the grand mean would also be shifted toward the center for the Likert-type scale. Finally, because the mean scores for the Likert-type scale are pushed toward the center of the distribution of organizations, the observed between-group variance for this scale will underestimate the true between-group variance.

There is an inherent problem with evaluating most of these potential effects: They necessarily involve comparisons of means and variances for different scales of measurement. That is, although theoretically the effect of Likert-type scaling should result in a relative inward shifting of the tails of the distribution, resulting in reduced between-group variance, the true scores are not on a Likert-type scale, and therefore these comparisons are not possible.¹ We can, however, observe the effect in a proportional index of variance such as ICC(1). Because the between-group variance will be reduced by the use of Likert-type scaling, the observed

proportion of between-group variance to total variance (i.e., ICC(1); Shrout & Fleiss, 1979) will underestimate the true proportion of between-group variance to total variance.

The influence on the reliability of the groups means (i.e., ICC(2)) is also likely to be affected. The influence on ICC(2), however, should be substantially reduced by another factor, namely group size. As Bliese (1998) has demonstrated, ICC(2) is greatly influenced by the average group size, and so the effect of reduced between-group variance discussed here is likely to be eliminated as group size increases. If so, we would expect to see an underestimation of ICC(2) primarily for very small group sizes, where the between-group variance component would weigh more heavily in estimation of ICC(2). Similarly, correlations between group-level aggregated variables and other group-level variables would also only be influenced by this effect when group sizes were very small. This hypothesis follows directly from our logic for the ICC(2) index and Bliese's finding that ICC(2), as an estimate of the reliability of group means, will determine the extent to which group-level correlations are attenuated.

Our final hypothesis concerns how these effects will change depending on the coarseness of the Likert-type scale. As fewer response options are used on the Likert-type scale, the effects on ICC(1) and ICC(2) will increase. The reasoning here is simple: As the number of response options decreases, the censoring of the distribution should increase. As such, all of the effects we described above should also increase as the number of response options decreases. There is a caveat to this notion, however. Two-point scales, despite suffering from a variety of other problems (Cohen, 1983), will not fall prey to the issue we discuss. Although a large portion of the true underlying scores will be censored on 2-point scales, they will not have the systematic effects described above because they cannot be pushed any further toward the middle of the distribution.

The Current Study

The current study sought to demonstrate these effects and examine the extent to which different formats of Likert-type scales moderate the effects. Prior research has examined how use of Likert-type scales on a single level of analysis results in increased measurement error as a result of loss of information between category thresholds (Glass, Peckham, & Sanders, 1972). Although this research suggests that most statistical tests are robust to this loss of information (Scheffé, 1959), we felt that the issue might be more important in a multilevel context for two reasons. First, as described above, multiple samples with means approaching the extreme ends of a scale will likely also suffer from censoring effects. Second, indices such as ICC(1) and ICC(2) may not be as robust to such measurement errors as analysis of variance.

In addition to the effects of Likert-type scaling, the amount of true variance both within and between groups should moderate these effects, resulting in greater underestimation of ICC(1), and, to a lesser extent, ICC(2) and group level correlations, as within- and between-group variance increases. To test these hypotheses, we conducted a series of Monte Carlo simulations, manipulating the number of scale points used to discretize the data and the amount of between-group variance and within-group variance. We began by examining situations where the grand mean fell in the middle of the response scale, but then we extended the simulations to examine what occurs when the grand mean drifts to either side of the response scale.

Method

Generating Multilevel Data

Simulations were conducted using a FORTRAN program running the International Mathematical and Statistical Libraries (IMSL). Two levels of data were created. First, the group-level data were generated by drawing a pseudo-random normal deviate using the RNNOF subroutine of IMSL. Thus, each group was first given a value for x_j that was a standard normal score with a mean of 0 and a SD of 1. To examine group-level correlations, a y_j score was generated ($M = 0$, $SD = 1$) and made to correlate with the group-level true x_j scores at a specified level for a specified number of groups (number of groups = 100, $\rho = .50$).²

For every x_j value, a number of within-group scores then were generated, also based on standard normal deviates ($M = 0$, $SD = 1$). These within-group scores then were added to the x_j value for each organization. The number of within-group scores was manipulated reflecting two group sample sizes ($n = 5$, $n = 50$).³ The resulting continuous individual-level scores (x_{ij}) therefore contained some amount of between-group variance and some amount of within-group variance.

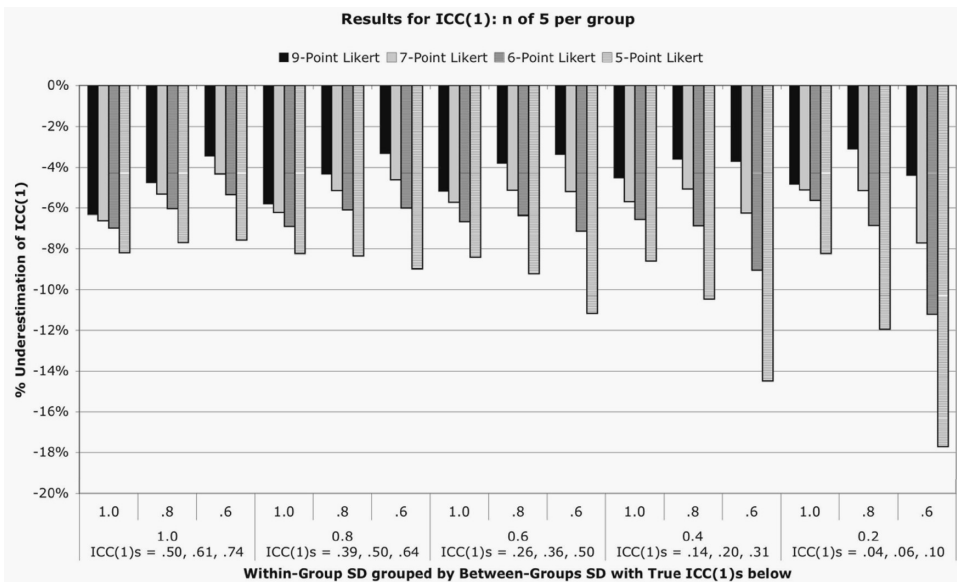
To manipulate within-group variance and between-group variance, constants were multiplied at each of the two steps discussed above. This procedure resulted in a manipulation of the SD of the continuous scores prior to being discretized into Likert-type scales. We selected three values for within-group SD (0.6, 0.8, and 1.0) and five values for between-group SD (0.2, 0.4, 0.6, 0.8, and 1.0) that appeared to generate total variance at realistic levels (total variance of continuous scores ranged from 0.4 to 2.0, and continuous data ICC(1)s ranged from approximately 0.04 to 0.74).

Likert-Type Scaling

In addition to the above continuous multilevel data, we also examined comparable data that had been discretized into Likert-type scales. To do this, each x_{ij} above was categorized using $p-1$ threshold values, where p is the number of response options in the Likert-type scale (e.g., a 5-point Likert-type scale used 4 threshold values). These threshold values were selected to create an even interval scale with extreme categories occurring at $-1.5 SD$ and $+1.5 SD$ from the mean. For example, a 5-point Likert-type scale had the values of -1.5 , -0.5 , 0.5 , and 1.5 for the 4 thresholds, whereas the 7-point Likert-type scale had the values of -1.5 , -0.9 , -0.3 , 0.3 , 0.9 , and 1.5 as its 6 thresholds. We decided to use this method because it resulted in a constant percentage of respondents in the extreme categories for each condition of within and between variance. Had we extended the outermost thresholds as the number of response options increased, our results for this factor would have been exaggerated. Thus, these thresholds were decided on because they are constant across the other conditions and they result in an approximately normal distribution when applied to standard normal deviates.

The above procedures were repeated 50,000 times for each condition. For each iteration, ICC(1) and ICC(2) were computed from the mean squares within groups and mean squares between groups, and correlations with the continuous y_j variable were computed both for continuous and Likert-type data. Results for each iteration were combined, and the average results across iterations are reported below.

Figure 1
Results for ICC(1): n of 5 Per Group



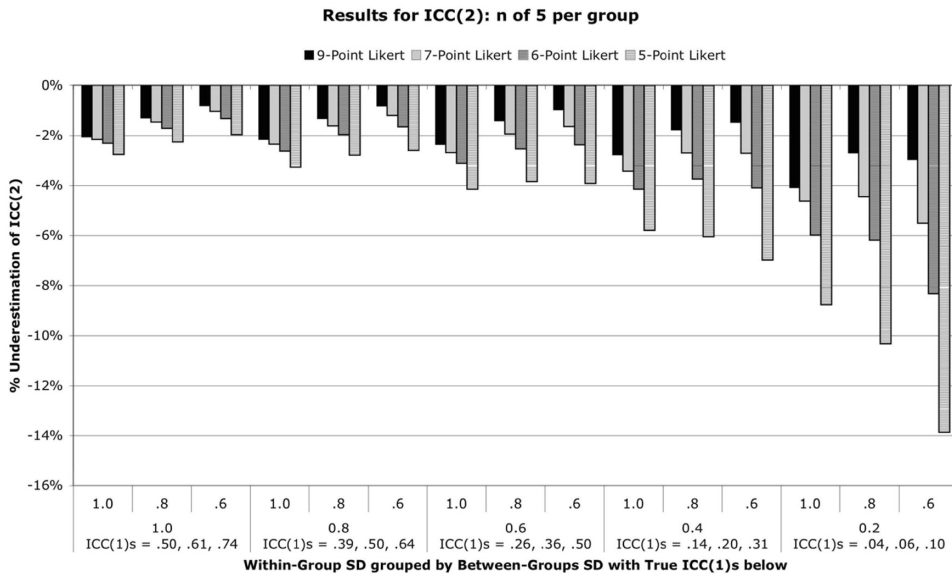
Results

Results for ICC(1)

Figure 1 depicts the results for underestimation of ICC(1) incorporating a combination of factors. First of all, the most interpretable pattern of results was obtained when underestimation of ICC(1) was examined as a proportional measure. That is, the estimates of ICC(1) using the various forms of Likert-type scales and between and within variances were not systematic when absolute deviations from the true ICC(1) were examined. As Figure 1 demonstrates, however, a definite pattern can be observed with the proportional underestimation of ICC(1). Consistent with our hypothesis, in all conditions, as the number of Likert-type scale response options decreased, the underestimation of ICC(1) increased.

The results for between- and within-group variance, however, exhibited a more complex interactive effect. At high levels of between-group variance, all Likert-type response formats increase in their underestimation of ICC(1) as within-group variance increases. As between-group variance decreases, however, this pattern disappears and then reverses, so that for the lowest levels of between-group variance, *decreases* in within-group variance actually create greater amounts of underestimation of ICC(1). It also appears that this tendency interacts with the number of Likert-type response options such that scales with fewer response options take on this reversal sooner than scales with more response options. For example, the underestimation for a 5-point Likert-type scale decreases with within-group variance only for the highest amount of between-group variance. A 9-point Likert-type scale, in contrast, exhibits this pattern for all but the lowest amount of between-group variance.

Figure 2
Results for ICC(2): n of 5 Per Group



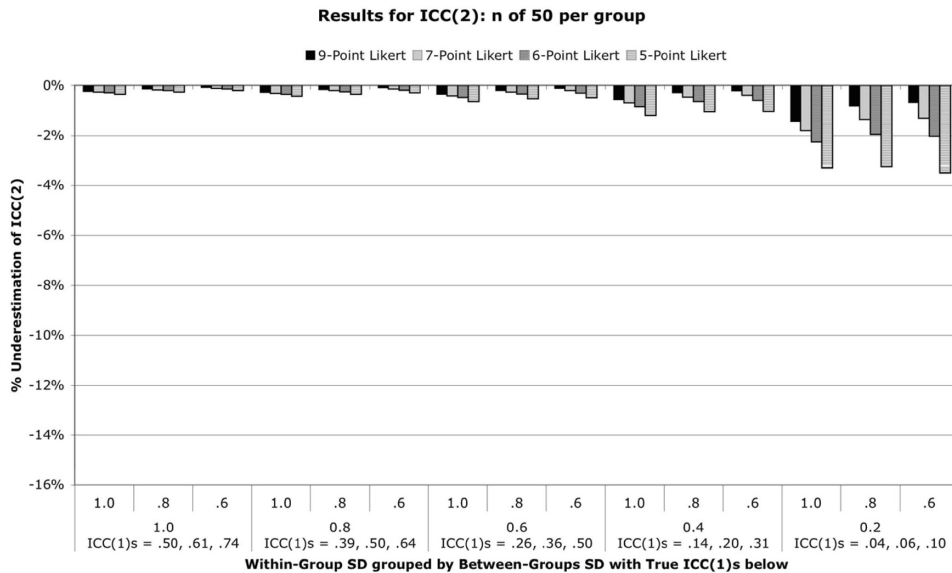
Thus, although our hypotheses appear to be supported for higher amounts of between-group variance, some additional factor seems to operate at lower levels of between-group variance. Note that we only present data here for one group size ($n = 5$). As we suspected, the pattern for ICC(1) was virtually identical for the larger group size ($n = 50$). In fact, the greatest discrepancy between all of the conditions for the $n = 5$ versus the $n = 50$ was a 00.8% difference. We note this to contrast it with the subsequent results for ICC(2).

We also have included the true ICC(1) values below the abscissa to facilitate computing the exact amounts of underestimation to aid interpretation. For example, the first condition on the far left of the graph involves a within-group SD of 1.0, a between-group SD of 1.0, and a 9-point Likert-type scale. The within- and between-group SD s correspond to a true ICC(1) of 0.50. Therefore, the observed 6% reduction (approximately) in the observed ICC(1) corresponds to a value of 0.47 (i.e., 0.50 minus 6% of 0.50).

Results for ICC(2)

Figures 2 and 3 present results for ICC(2) for group sizes of 5 and 50, respectively. Again, a consistent pattern was observed only when results were presented as the percentage underestimation of ICC(2). As can be seen, the pattern for ICC(2) for a group size of 5 is almost identical to that of ICC(1), although the absolute magnitude of the underestimation is less severe. Again, as the number of Likert-type scale response options decreased, the underestimation of ICC(2) increased in all conditions. The interactive pattern among number of response options, between-group variance, and within-group variance is exactly

Figure 3
Results for ICC(2): n of 50 Per Group



the same as that reported above for ICC(1). The primary difference between the results for ICC(1) and ICC(2) occurs when group size increases. As Figure 3 demonstrates, as group size increases, the underestimation of ICC(2) is virtually eliminated, although the pattern of effects remains the same.

Results for r_{xy}

The results for group-level correlations displayed in Figures 4 and 5 did not display the complex interactive pattern observed for ICC(1) and ICC(2). The effect for number of Likert-type scale response options was present in all conditions but was not as noticeable as was the case for ICC(1) and ICC(2). The pattern associated with level of between- and within-group variance is consistent with the effects described in some detail by Bliese (1998) and, as suspected, all but disappears when group size increases. These results replicate Bliese's finding of the massive amounts of attenuation that occur with small group sizes, low ICC(1) values, and low ICC(2) values.

Results for Shifted Grand Mean

In all of the results reported so far, the grand mean of the continuous and Likert-type scales was in the middle of the scale. In realistic situations, however, such a perfect balance of responses may not be terribly common. Therefore, we sought to determine the same effects on ICC(1), ICC(2), and r_{xy} when the grand mean approached one or the other end of

Figure 4
Results for r : n of 5 Per Group

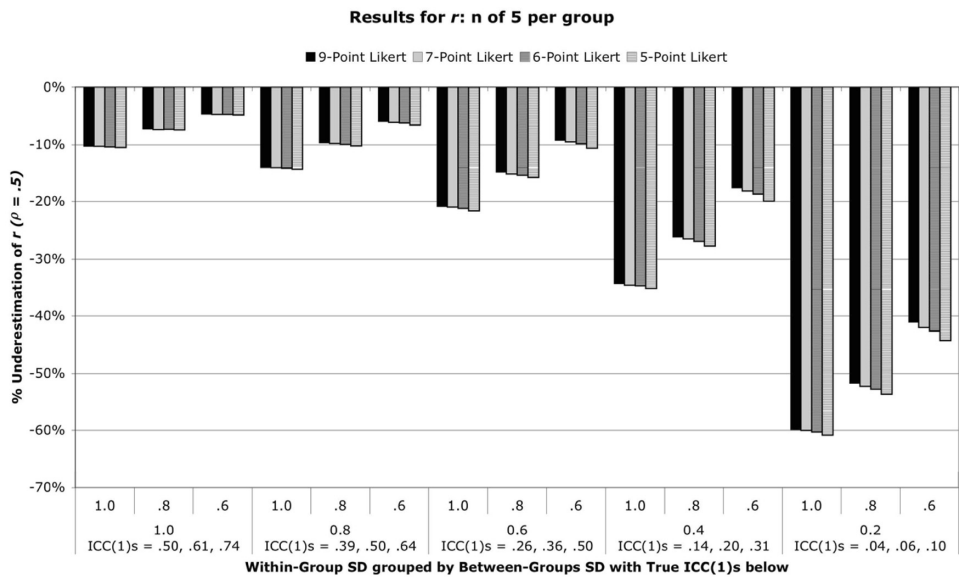


Figure 5
Results for r : n of 50 Per Group

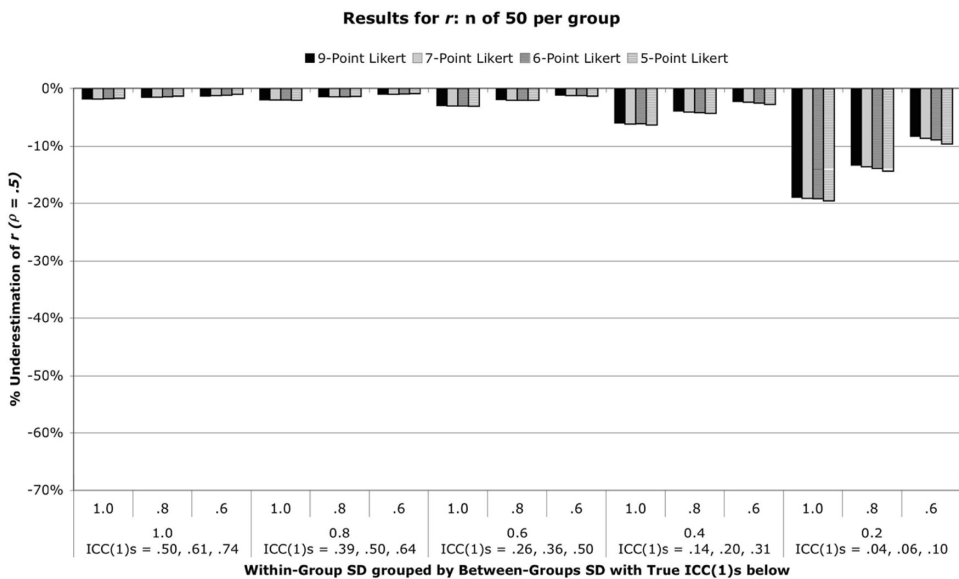
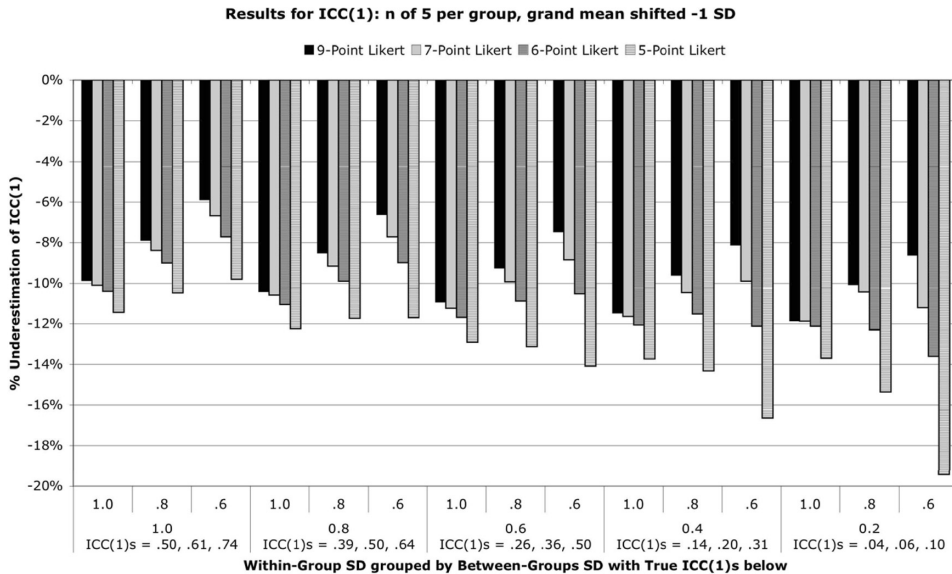


Figure 6
Results for ICC(1): n of 5 Per Group, Grand Mean Shifted -1 SD



the continuum. Our expectations were straightforward: As the grand mean deviates from the middle of the distribution, an increasing amount of the scores would be censored. Therefore, all of the effects described above should become more severe. The results revealed exactly this pattern. Figure 6 demonstrates the magnitude of these effects for ICC(1) when the grand mean of the continuous distribution was shifted 1 SD in the negative direction.⁴ The only other noteworthy effect of shifting the grand mean was that it tended to reduce the reversal pattern discussed above with respect to decreasing between-group variance. Although we only present a figure for ICC(1), the pattern of results was similar for the other indices.

Discussion

The effects of Likert-type scaling appear to have a strong influence on the interpretation of some aspects of group scores but less of an influence on others. For ICC(1), using Likert-type scales resulted in systematic underestimation of the true proportion of between-group variance. This effect was consistent across all conditions and increased as fewer response options were available on the Likert-type scale. Although the magnitude of the underestimation ranged from 3% to almost 20%, the higher levels of underestimation were not reserved for just the more extreme conditions. For example, even when the grand mean is exactly in the middle of the scale and the amounts of between- and within-group variance reflect realistic data conditions (e.g., continuous score SD s of 0.4 and 0.8, respectively), use of a 5-point

Likert-type scale still resulted in underestimating the true proportion of between-group variance by 10%. These findings are consistent with our explanation of the censoring that occurs for groups whose true scores fall in the tails of the between-group distribution.

ICC(2) reflected a similar pattern of effects, but the underestimation was less severe, ranging across all conditions from less than 1% to more than 15%. Again, the effect varied considerably with the number of response options present in the Likert-type scale, with greater levels of underestimation occurring with fewer response options. Unlike the results for ICC(1), but as we anticipated, the effects were substantially reduced as group size increased. In fact, applying the Spearman-Brown prophecy formula to our obtained underestimated ICC(1) coefficients (see Bliese, 1998) reproduced the effects found for ICC(2) very closely for the large group size condition.⁵

The effects on both ICC(1) and ICC(2) for within- and between-group variance were not quite as straightforward in their interpretation. Although increases in within-group variance did occasionally yield greater underestimation of ICC(1) and ICC(2), this pattern held only for situations where between-group variance was also high. It is interesting that at lower levels of between-group variance, this pattern did not simply disappear; rather, it completely reversed, such that greater underestimation of ICC(1) and ICC(2) occurred as within-group variance decreased.

Although we did not predict this reversal to occur, we can, in hindsight, offer a potential explanation. Because the effect reversed when between-group variance was very low, most responses that occur when within-variance is also low (particularly for scales with fewer response options) may fall mostly into one or two categories. Discrepancies from these categories are rare, but when they do occur, they serve to overestimate the within-group variance. Between-group variance is less subject to this effect as it is not constrained to one or two discrete values. The end result is that within-group variance is overestimated more so than between-group variance, yielding underestimates of ICC(1) and ICC(2). As true within-group variance increases, it is likely that more than one or two response options will be selected and that the reversal pattern will dissipate, giving way for the effects due to censoring at the extreme ends.

So it appears that the effects of censoring apply when there are relatively greater amounts of variance, but a completely different effect occurs as variance is constrained. Interestingly, this second effect hinges even more strongly on the number of points in the Likert-type scale. Also of note is that the result of either effect is the same: an increasing underestimation of ICC(1), contributing, particularly with small group sizes, to unreliability of the group means.

Although underestimation of the true proportion of between-group variance and increased unreliability of group means are issues of concern in their own right, it was somewhat relieving to observe that Likert-type scale usage has little ultimate effect on the magnitude of group-level correlations. At first blush, however, it may not appear that our conditions had little effect on the group-level correlations. Indeed, Figure 4 depicts up to a 60% reduction in the group-level correlation. It is important to point out, however, that these reductions are due in large part to the known combination of effects of group size and ICC(1) on group-level correlations (Bliese, 1998). That is, by manipulating between- and within-group variance, we created situations where the true proportion of between-group variance (i.e., true ICC(1)) was low. For a given group size, lower true ICC(1)s result in unreliable

group means, and this unreliability has a dramatic effect on the group-level correlation. By comparison, the effects of Likert-type scaling on group-level correlations are relatively minor.

Implications for Organizational Researchers

To understand the implications of our findings for organizational researchers, we must consider how and when we are most interested in ICC(1) and ICC(2). For example, quite often, ICC(1) is used to evaluate the presence of group-level effects that may exist beyond individual-level effects (Bliese, 1998). The results obtained here suggest that when Likert-type scales are used, researchers might overestimate the extent to which a relation is occurring at the individual-level if they base such decisions on an evaluation of ICCs. For example, a researcher may have a theoretical interest in interpreting a group-level effect, but a low ICC(1) value brings into question such an interpretation. This researcher may wish to consider the scale format of the lower level data and modify his or her interpretation appropriately. In contrast, one can imagine using ICC(1) as evidence of a potential frog pond effect (Bliese, 2000). If the use of Likert-type scales creates underestimates of ICC(1), then the researcher might not be justified in leaping to this conclusion. Note also that these results extend to other multilevel situations. For example, if researchers are examining people over time and use Likert-type scales at each time point, then the coarseness of the scale could contribute to the conclusion that effects are primarily within person in nature when in fact a larger person-level effect is occurring.

Similarly, underestimating ICC(2) can result in inappropriate conclusions concerning whether the groups' mean scores are reliable enough to be used for estimation of group-level effects. Ironically, if researchers using Likert-type scales and aggregate variables use ICC(2) to disattenuate group-level correlations (Bliese, 1998), then they are likely to be overestimating, albeit slightly, the true group-level correlation (i.e., because the ICC(2) will be an *underestimate* of true group mean reliability). Recall, however, that our simulation results apply when only one of the two covariates is in the Likert-type format. The overestimation of the disattenuated group-level correlation may not be so slight if both variables are Likert-type variables.

Another source of irony involving these results surfaces for researchers who wish to escape the measurement errors that occur at an individual level by simply aggregating variables to the group level. Ecological fallacies notwithstanding, these results demonstrate that some aspects of measurement error (i.e., using scales that do not adequately represent the underlying continuous distribution) will not only be carried forward into the aggregate variable, but might also have an even larger effect at the aggregate level than at the individual level.

Importantly, the effects uncovered in these simulations are eliminated fairly easily: If Likert-type scales are to be used, response formats with a larger number of options (e.g., 7 points or 9 points) are advisable. This recommendation is also consistent with the recommendations of researchers examining number of response options in single-level research (e.g., Cicchetti et al., 1985). The possibility of censored data is of course present within a single level of analysis, but we argue that it becomes more of an issue in multilevel data sets as the between-group variance creates greater opportunities for such censoring to occur.

Limitations

A single simulation study can never explore all of the idiosyncrasies that occur in actual multilevel data sets. As such, there are undoubtedly numerous factors that convene to reduce, eliminate, or expand the problems discussed here. For example, the current simulations examined what occurs when examining a single-item scale. It is not clear how these results would change for multi-item scales, but several possibilities can be suggested. We suspect that the reversal pattern that occurs for lower levels of between- and within-group variance would be greatly reduced. That is, if our current hypothesis, that within-group scores are largely constrained in these conditions to one or two values, creating inaccurate overestimates of within-group variance, is correct, then adding multiple items to the scale should reduce or eliminate the inaccuracy yielding estimates of ICCs that are closer to their true values.

The results of using multi-item scales are less clear when considering the effects of censoring. For example, if a group has a true score that is 2 *SDs* below the grand mean, then the current study demonstrates that a Likert-type scale is likely to result in a censoring effect that shifts this group's mean above its true $-2 z$ score. If multiple items were used, then it seems possible that the same amount of censoring would occur, on average, for each item in the scale. So the same issue described in our article would occur for the items that went into the scale mean. Indeed, although we have framed our discussion in terms of individuals nested within groups, the same issues should occur for items nested within people. That is, one could interpret our results as an examination of scale means (as opposed to group means) consisting of 5 or 50 Likert-type scale items (vs. continuously scaled items) across differing true levels of item variance and scale variance. If viewed in this manner, then our results suggest that multi-item scales will consist of inwardly biased scale means. Thus, if one were to use multi-item scales, these inwardly biased scale means would be carried forth into the group means, effectively shifting the same problem to a lower level of analysis. Future research must determine whether this process would exacerbate or ameliorate the results for higher levels.

Another limitation is that we did not examine methods by which such censoring effects could be reduced or eliminated. For example, researchers in economics have often employed robust regression techniques in situations similar to the ones described here (Maddala, 1983). In addition, current statistical software packages now allow for the implementation of such techniques within a more general structural equation modeling framework (Muthén, 1989). By specifying an appropriate model (i.e., one with censoring that occurs at the floor and ceiling of a scale), it might be possible to reduce or eliminate the underestimates of ICCs observed here.

Future research may answer these and other questions concerning how violations of statistical assumptions at one level could be magnified at a higher level of analysis. The results here offer a cautionary note to researchers using Likert-type scales with relatively few response options and shed light into how such scales operate at an aggregate level.

Notes

1. Typically, when one wishes to compare scores on different scales of measurement, some form of standardization (e.g., *z* scores) will help. Note, however, that in this case, we wish to compare changes in variance across two different scales. Standardization would result in equating of the variances, making comparisons of this sort impossible.

2. We chose the values of 100 and 0.5 to simplify presentation and to make our results comparable with those of Bliese (1998). We note that simulations using other population correlations and other numbers of groups obtained comparable results.

3. We suspected that the influence of using Likert-type scales on ICC(2) would depend greatly on group size (Bliese, 1998), with effects primarily being observed for small group sizes. Although we conducted simulations for a wide variety of group sizes, for the purposes of examining impact on ICC(2), we have selected a subset that best illustrates these effects.

4. We report only the results for a shift in the negative direction. We could think of no reason to believe different effects would occur with a shift in the positive direction. Indeed, further simulation results verified completely symmetric effects for positive shifts.

5. Bliese (1998) points out that such estimation becomes more accurate as group size increases; thus, our estimation was extremely accurate (within 1% of the observed ICC(2) value) for the group size of 50 but is variably accurate (up to 12% higher) when the group size is 5.

References

- 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 analyses. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco: Jossey-Bass.
- Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*, 234-246.
- Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods, 8*, 375-409.
- Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Journal of Applied Psychology, 9*, 31-36.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.
- Gangestad, S., & Snyder, M. (2000). Self-monitoring: Appraisal and reappraisal. *Psychological Bulletin, 126*, 530-555.
- George, J. M., & James, L. R. (1993). Personality, affect, and behavior in groups revisited: Comment on aggregation, levels of analysis, and a recent application of within and between analysis. *Journal of Applied Psychology, 78*, 798-804.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variances and covariances. *Review of Educational Research, 42*, 237-288.
- Gupta, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika, 39*, 260-273.
- Hofmann, S. A., Griffin, M. A., & Gavin, M. B. (2000). The application of hierarchical linear modeling to organizational research. In K. J. Klein & S. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 467-511). San Francisco: Jossey-Bass.
- Hox, J. J. (2002). *Multilevel analysis*. Mahwah, NJ: Lawrence Erlbaum.
- James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*, 219-229.
- Jones, A. P., & James, L. R. (1979). Psychological climate: Dimensions and relationships of individual and aggregated work environment perceptions. *Organizational Behavior and Human Performance, 23*, 201-250.
- Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational & Psychological Measurement, 4*, 987-995.
- Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions*. New York: Wiley-Interscience.
- Krieg, E. F. (1999). Biases induced by course measurement scales. *Educational & Psychological Measurement, 59*, 749-766.

- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- 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.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- Micceri, T. (1989). The Unicorn, the normal curve and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Muthén, B. (1989). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology*, 42, 241-250.
- Pearson, K., & Lee, A. (1908). On the generalized probable error in the multiple normal correlation. *Biometrika*, 6, 59-68.
- Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin*, 105, 167-170.
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.
- Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology*, 77, 336-342.
- Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley.
- Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.

Daniel J. Beal is an assistant professor of psychology at Rice University. He received his PhD in psychology from Tulane University. His research interests include within-person processes linking affect and performance and several methodological topics, including multilevel and longitudinal modeling and meta-analytic techniques.

Jeremy Dawson is a research fellow at Aston Business School, Aston University, England. His research interests include measurement of group diversity, moderated multiple regression, and analysis of incomplete data in group studies. His publications have appeared in journals such as the *Journal of Applied Psychology* and the *Leadership Quarterly*.