

RESEARCH REPORT

Improving the Measurement of Group-Level Constructs by Optimizing
Between-Group DifferentiationPaul D. Bliese, Mark A. Maltarich,
and Jonathan L. Hendricks
University of South CarolinaDavid A. Hofmann
The University of North Carolina at Chapel HillAmy B. Adler
Walter Reed Army Institute of Research, Silver Spring, Maryland

The ability to detect differences between groups partially impacts how useful a group-level variable will be for subsequent analyses. Direct consensus and referent-shift consensus group-level constructs are often measured by aggregating group member responses to multi-item scales. We show that current measurement validation practice for these group-level constructs may not be optimized with respect to differentiating groups. More specifically, a 10-year review of multilevel articles in top journals reveals that multilevel measurement validation primarily relies on procedures designed for individual-level constructs. These procedures likely miss important information about how well each specific scale item differentiates between groups. We propose that group-level measurement validation be augmented with information about each scale item's ability to differentiate groups. Using previously published datasets, we demonstrate how ICC(1) estimates for each item of a scale provide unique information and can produce group-level scales with higher ICC(1) values that enhance predictive validity. We recommend that researchers supplement conventional measurement validation information with information about item-level ICC(1) values when developing or modifying scales to assess group-level constructs.

Keywords: multilevel, reliability, validity, measurement

Group-level attributes can be assessed using expert ratings (Hirschfeld, Jordan, Feild, Giles, & Armenakis, 2006), group discussion (Gibson, Randel, & Earley, 2000), minimum or maximum values (Barrick, Stewart, Neubert, & Mount, 1998), or variance-based dis-

person (see Kozlowski & Klein, 2000). Often, however, group-level attributes are assessed by averaging group members' responses to scores from multi-item scales. This latter form of measurement underlies "direct consensus" or "referent-shift consensus" constructs (Chan, 1998), and has been used to assess group attributes such as collective efficacy, cohesion, safety climate, and leadership (e.g., Bandura, 2000; G. Chen & Bliese, 2002; Hofmann & Mark, 2006; Mathieu, Kukenberger, D'Innocenzo, & Reilly, 2015; Tasa, Taggar, & Seijts, 2007; Zohar, 1980).

Regardless of how group attributes are assessed, the subsequent utility of any group-level measure depends, in large part, on how well the measure detects group differences: Measures that fail to detect group differences are generally of low practical or research value. We show that measurement practices used over the last 10 years may not optimize group differentiation, and we propose a simple modification that can be integrated with existing best practices in multilevel scale validation (e.g., G. Chen, Mathieu, & Bliese, 2004) to enhance the measurement properties of scales developed to be used as group-level measures.

Group Differentiation and Review of Practice

Detecting Group-Level Relationships

To set the stage for understanding the measurement significance of between-groups differentiation, we review two common multi-

Editor's Note. Vicente Gonzalez-Roma served as the action editor for this article.—GC

This article was published Online First September 17, 2018.

Paul D. Bliese, Mark A. Maltarich, and Jonathan L. Hendricks, Management Department, Darla Moore School of Business, University of South Carolina; David A. Hofmann, Kenan-Flagler Business School, The University of North Carolina at Chapel Hill; Amy B. Adler, Walter Reed Army Institute of Research, Silver Spring, Maryland.

Paul D. Bliese is also affiliated with the Walter Reed Army Institute of Research, Silver Spring, Maryland.

Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Department of the Army or the Department of Defense. We also want to thank Barbara Mark for providing access to the data used in the third example.

Correspondence concerning this article should be addressed to Paul D. Bliese, Management Department, Darla Moore School of Business, University of South Carolina, 1014 Greene Street, Columbia, SC 29208. E-mail: paul.bliese@moore.sc.edu

level metrics—the ICC(1) and the ICC(2)—and show how these indices are related to the magnitude of group-level correlations. The ICC(1) represents the amount of variance in any one group member's response that can be explained by group membership, and the ICC(2) represents an estimate of the reliability of the group mean (see Bliese, 2000; LeBreton & Senter, 2008). The ICC(2) is a function of group size and the ICC(1) and can be estimated using the Spearman-Brown reliability formula (Shrout & Fleiss, 1979).

Importance of ICC(2). The magnitude of the ICC(2) plays a critical role in the strength of correlations between group-level variables (Bliese, 2000; LeBreton & Senter, 2008; Geldhof, Preacher, & Zyphur, 2014; Raudenbush, Rowan, & Kang, 1991). For example, consider individual ratings of (a) task significance and (b) anger/hostility in an illustrative dataset containing 2,042 soldiers nested in 49 U.S. Army companies (e.g., Bliese, Halverson, & Schriesheim, 2002; Cohen, Doveh, & Nahum-Shani, 2009; Doran, Bates, Bliese, & Dowling, 2007). Observed ICC(1) values are 0.081 and 0.055 for task significance and anger, respectively, and ICC(2) values are 0.786 and 0.709. The group-level correlation between task significance and anger ($r = -0.697$) differs from the within-group correlation ($r = -0.312$), suggesting an emergent group-level relationship.¹

Randomly selecting 500 of the 2,042 observations and estimating ICCs and correlations 1,000 times returns ICC(1) values that are functionally unchanged (average value of 0.082 for task significance and 0.052 for anger). The within-correlation is likewise unchanged (average value of -0.315). As expected, however, the ICC(2) drops from 0.786 to 0.465 for task significance and from 0.709 to 0.343 for anger. Similarly, the between-groups correlation is now -0.521 (vs. -0.697), demonstrating the ICC(2)'s role in the magnitude of group-level relationships. As another example, we can randomly assign group membership to each participant, thereby eliminating any meaningful between-groups differentiation while keeping group size effects constant. The resulting ICC(1) average values for both task significance and hostility over 1,000 trials are both .002. The ICC(2) values are .056 and .061 for task significance and hostility, respectively.² Under these conditions, the raw correlation, within-correlation, and group-level correlation are functionally equal (-0.345 , -0.351 , and -0.345 , respectively). Other examples showing the importance of the ICC(2) to group-level correlations in simulated and actual data are provided by Bliese, Maltarich, and Hendricks (2018).³

Scale development, ICC(2), and ICC(1). Reasonably large ICC(2) values can be obtained by ensuring that group sizes are large (as long as ICC(1) values are nonzero); however, if group sizes are fixed, the only way to enhance a scale's ICC(2) value is to optimize ICC(1). Our goal is to provide practical advice on how to optimize ICC(1) values of multi-item scales by illustrating how each item underlying a scale score can differ with respect to ICC(1) values in ways that are not detected in typical measurement validation.

We integrate our ideas into G. Chen et al.'s (2004) influential five-step framework for multilevel construct validation. As Chen et al. noted, "multi-level construct validation is substantially more complex than its single-level analogue" (G. Chen et al., 2004, p. 275). We augment G. Chen et al. (2004) by showing the relevance of considering scale item ICC(1) values in the early phases of construct validation (see also Tay, Woo, & Vermunt, 2014). We emphasize that the process of establishing valid measures is mul-

tifaceted, and a scale's ability to differentiate groups is but one part of the process; nonetheless, we demonstrate how examining ICC(1) estimates for each item of a scale can enhance the ICC(1) values of the overall scale.⁴ Finally, we further illustrate how maximizing group-level properties of a scale facilitates detecting aggregate relationships.

Multilevel Measurement Validation and Current Practice

As detailed in G. Chen et al. (2004), the measurement validation of group-level constructs, like the measurement of individual-level constructs (Hinkin, 1995, 1998; Schwab, 1980), relies on establishing validity and reliability to ensure the intended construct is being assessed with minimal error. To this end, G. Chen et al. (2004) present a five-step process: (1) defining the construct, (2) articulating the nature of the aggregate construct (e.g., consensus, referent shift, etc.), (3) examining the psychometric properties of constructs across levels, (4) examining construct variability between units, and (5) examining construct function across levels.

The core ideas proposed by G. Chen et al. (2004) provide a foundation for developing group-level measures; however, we expand on ideas presented in Step 4 (examining construct variability between units) by showing that group-level variability is partially influenced by decisions made in Step 3 (examining psychometric properties across levels). The idea that scale items might differentially reflect group-level properties builds on a body of work advocating that researchers consider the unit of analysis when testing psychometric properties (Pornprasertmanit, Lee, & Preacher, 2014; Tay et al., 2014; Zyphur, Kaplan, & Christian, 2008). Despite this work, we show that current measurement validation practices typically rely on individual-level properties of scale items—a practice that overlooks an item's ability to differentiate groups.

To examine current practice, we searched articles in the *Academy of Management Journal*, *Journal of Applied Psychology*, and the *Journal of Occupational Health Psychology* that included the keywords "multilevel," "scale," "validation," or "climate" between 2008 and 2017. Our initial search identified 456 articles. We filtered our selection to only include articles that measured a unitary collective property by surveying group members with scales of direct consensus or referent shift items. The criteria yielded 133 studies.

¹ Our point here is not to develop a theoretical framework around emergent properties, but the results certainly suggest on both a practical and theoretical level that working in a group that collectively does or does not view task significance highly has implications for the average anger and hostility in the group that go beyond understanding how individual views of task significance relate to individual anger and hostility.

² There is a slight upward bias because these values are estimated from a random effects model where negative values are not possible (variances cannot be less than zero).

³ Bliese et al. (2018) further illustrate that emergent effects involving group means can be detected when ICC(2) values are well below norms of 0.70, but the key point to the current article is that ICC(2) values are related to the ability to detect meaningful group-level correlations.

⁴ We focus on the ICC(1) from this point on to keep a metric invariant to group-size differences. Ultimately, however, our interest in ICC(1) links to its ability to detect group-level correlations via the ICC(2).

The results of our review are summarized in Table 1. The most common practice for assessing the reliability of group-level scales was to report Cronbach's alpha on individual-level means, although 15 articles (11%) reported at least one alpha based on the group means. The basis of the alpha calculation was unclear for at least one of the constructs in 35 articles (26%). After establishing individual-level reliability, researchers typically justified aggregating constructs to the group level, appealing to ICC(1) (117 articles, 88%), ICC(2) (95 articles, 71%), $r_{wg(j)}$ (98 articles, 74%), or some combination. All three aggregation measures were reported for at least one construct in 82 (62%) articles. Only 14 studies (11%) leveraged between-groups information from multilevel confirmatory factor analysis (ML-CFA). Table 1 also provides information about item referents. In total, 72 (54%) of the articles used only referent shift items. The remaining articles used either nonreferent shift or a mix.⁵

Studies where authors removed items from group-level constructs (15 studies, or 11%) are particularly relevant. We summarize authors' justifications in Table 2. Seven studies dropped items based on individual-level alpha, factor analysis, or inter-item correlations (e.g., Nishii, 2013; Sliter, 2013). Eight studies reduced items on the basis of their conceptual content, including two studies that dropped negatively worded items (e.g., G. Chen, Farh, Campbell-Bush, Wu, & Wu, 2013; Priesemuth, Schminke, Ambrose, & Folger, 2014), and five studies dropped items for low content validity (e.g., Aryee, Walumbwa, Seidu, & Otaye, 2012; Z. Chen, Zhu, & Zhou, 2015). Two studies (Hammer, Kossek, Bodner, & Crain, 2013; Mikolon, Kreiner, & Wieseke, 2016) reported item-level ICCs, but neither study used that information for decisions about dropping items. Several articles used individual-level psychometric properties in item selection and then validated collective constructs based on the reduced scales (Aryee et al., 2012; Scott & Judge, 2009; Zweber, Henning, & Magley, 2016).

We conclude that scale development and multilevel validation are generally conceived separately; perhaps the best example of this is found in a study conducted by X. P. Chen, Liu, and Portnoy

Table 2

Reported Criteria for Item selection in Studies With Shortened Scales

Criterion for item selection	Articles reporting (Count)	Articles reporting (Percent)
Individual relationships	7	47%
Conceptual Content	7	47%
CFA from prior research	1	7%

Note. $N = 15$ articles that used shortened scales. Individual relationships include inter-item correlations and factor loadings from confirmatory or exploratory factor analysis.

(2012). The authors were thorough in their analyses. They reported ICC and $r_{wg(j)}$ values to justify aggregation, and conducted an ML-CFA. For scale development, however, they used only individual-level psychometric information. Specifically, they trimmed their scale on the basis of individual-level factor loadings before assessing its multilevel properties. We emphasize that we are not critical of X. P. Chen et al. (2012), because understanding patterns of relationships shown by individual psychometric tests is important to scale development. Rather, our goal is to consider how to improve the process by further integrating information that helps differentiate groups.

Group-Level Properties of Scale Items

Using large samples, we demonstrate that scale items provide unique information that can enhance the ability to detect group differences. We illustrate our points with three examples. First, we examine items used to measure leadership and cohesion in a large U.S. Army sample (Adler et al., 2015; Foran & Adler, 2013). The leadership example contains a mix of items referencing the respondent, the group, and the leader, and we can make an *a priori* prediction that items referencing the respondent will perform least well in terms of differentiating groups based on findings by Klein, Conn, Smith, and Sorra (2001), who found items with an individual referent (vs. group) displayed significantly lower between-group variability (see also, Arthur, Bell, & Edwards, 2007). In our leadership example, we show that a comprehensive set of individual-based psychometric tests fails to capture information evident in the ICC(1) values. Our second example shows that the group-level properties of three cohesion items are temporally rank-order stable, and further illustrates that item ICC(1) values are largely unaffected by unequal group sizes. In our third example, we examine a measure of safety climate within a hospital setting to test the predictive validity of shortened scales based on different criteria. Our examples represent common uses of multilevel constructs across different measures and settings.

Example 1: Drill Sergeant Leadership

Example 1 involves a subset of items used by new U.S. Army recruits to assess their drill sergeants. The complete 18-item drill sergeant scale was used as a control in a randomized trial of sports

Table 1

Review of Reported Individual- and Group-Level Measures for Group Constructs

Scale assessment	Articles reporting (Count)	Articles reporting (Percent)
Individual alpha	84	63%
Group alpha	15	11%
Ambiguous alpha	35	26%
No alpha reported	6	5%
ICC1	117	88%
ICC2	95	71%
$r_{wg(j)}$	98	74%
ICC1, ICC2, & $r_{wg(j)}$	82	62%
Only non-referent shift	31	23%
Only referent shift	72	54%
Both referent & non-referent shift	30	23%

Note. $N = 133$ articles. Ambiguous alpha refers to instances in which it is unclear whether the alpha level was at the individual or group level. The percent of studies reporting ICC and $r_{wg(j)}$ values sum to more than one because many studies reported more than one of these values. Similarly, several articles reported both individual and group alphas. Referent shift tabulations include only scales where the approach was clear.

⁵ Results were generally similar when summarized at the scale level: 203 of the 323 (63%) scales where we could make a determination used referent-shift items, and 37% (120) used direct consensus.

psychology self-talk skills (Adler et al., 2015). Using another large sample, Foran and Adler (2013) performed a detailed analysis of the individual-level psychometrics of the 18-item scale and determined that the items loaded on three factors, “motivate,” “respect,” and “tough.” We use the Adler et al. (2015) sample to examine the group-level properties of the 9 items in the “motivate” category. The sample contains data from 2,199 soldiers in 47 platoons. Group sizes ranged from 37 to 58 ($M = 46.70$, $SD = 3.98$).

We emphasize that we are not attempting to validate a new scale. Rather, we show that if we wanted to further optimize a group-level “motivate” drill sergeant scale by enhancing its ability to detect platoon differences, we would gain unique information by examining item-level ICC(1) values. Items differ in one important way: Some focus on the respondent (“Push me to do my best”); other items focus on the group (“Give up on us”) or leaders (“Are motivated”).

Table 3 provides the overall alpha (top), item total correlations, and alpha values without each item for the raw (Total), group-mean centered (Within), and group-mean (Between) variants of the measure. The last 3 items (items 7, 8 and 9) are reverse-coded and generally have the lowest item-total correlations. These three would be candidates for removal if one wanted to shorten the scale, but the reliabilities and item-total correlations are reasonable. Table 4 provides item loadings from three variants of confirmatory factor analyses tests: a raw CFA (CFA Total), a CFA based on group-mean centered responses (CFA Within), a CFA based on group means (CFA Between), and an ML-CFA that simultaneously estimates Within and Between loadings. The CFA Total ($\chi^2 = 875.48$; $p < .001$; $CFI = 0.94$; $SRMR = 0.04$), CFA Within ($\chi^2 = 716.36$; $p < .001$; $CFI = 0.94$; $SRMR = 0.04$), CFA Between ($\chi^2 = 63.88$; $p < .001$; $CFI = 0.95$; $SRMR = 0.02$), and ML-CFA ($\chi^2 = 749.33$; $p < .001$; $CFI = 0.94$; $SRMR_{within} = 0.04$; $SRMR_{between} = 0.02$) suggest reasonable fit. With the exception of ML-CFA Between, the last three item loadings generally mirror the item-total correlations by having the lowest values (item 8 slightly exceeds item 1 in the CFA Between and item 7 exceeds items 1 and 2 in the ML-CFA Within).

The last column in Table 4 provides the ICC(1) estimates with approximate 95% confidence intervals. Note the large differences in ICC(1) values across items. As anticipated, items with the individual as the referent (items 1 and 3–7) tend to have low values. ICC(1) values tend to be similar to the ML-CFA-Between factor loadings standardized on Within and Between variances (Heck, 2001); however, the ML-CFA Between results and the ICC(1) results do not provide entirely redundant information. For instance, item 1 has a loading of .45 and item 4 has a loading of .46; nonetheless, the ICC(1) value is higher for item 1 (0.10) than for item 4 (0.08).

One could imagine using these results to inform the selection of items for a scale. Using individual-level results, the scale might be reduced by eliminating the last three reverse-coded items. The ICC(1) for the scale based on 9 items is .18, but the elimination of the last three items reduces the ICC(1) to .14 (roughly a 25% decrease in the scale ICC(1) value). An alternative strategy might be to eliminate items based on ICC(1) or ML-CFA-Between results. Eliminating the 3 items with the lowest ICC(1) produces a scale with an ICC(1) of .22, and eliminating the 3 items with the lowest ML-CFA-Between loadings produces a scale with an ICC(1) of .21. Both ap-

proaches increase the scale’s ability to differentiate groups. The key implication is that it may important to keep items 8 and 9 even though these items have some of the weaker item-total correlations and factor loadings in the individual-level tests. The ICC(1) tests provide unique information relevant to measuring drill sergeant differences across platoons.

Example 2: Cohesion

Our second example focuses on the stability of ICC(1) values over three time periods, and allows us to examine the impact of unbalanced group sizes and further examine how variations in group size are related to item-level ICC(1) values. We expect some minor change in ICC(1) values over time for two reasons. First, each ICC(1) value represents a point-estimate, so values will vary by chance alone. Second, it is likely that responses will become more consistent within groups and/or diverge across groups, producing changes in ICC(1) values (see Lang, Bliese, & de Voogt, 2018). Nonetheless, we expect rank order stability of item ICC(1) values.

We use the same Adler et al. (2015) sample, but focus on ratings of group cohesiveness assessed using 3 of the 6 items reported by Podsakoff and MacKenzie (1994; see also Podsakoff, MacKenzie, & Fetter, 1993; Podsakoff, Niehoff, MacKenzie, & Williams, 1993).⁶ The 3 items are listed in Table 5. The scale has favorable psychometric properties and ICC values (Chiniara & Bentein, 2016; Podsakoff, MacKenzie, & Ahearne, 1997; Williams et al., 2016).

Item ICC(1) values over the three time intervals ranged from .11 to .20 (see Table 5). Notice that the magnitudes change, but the rank order remains unchanged: Item 1 consistently has the highest ICC(1) value, and Item 3 consistently has the lowest value. To illustrate that results are robust to unbalanced data (i.e., differences in group sizes), we held all group sizes constant at 27 (the smallest group size in one group at Time 3) by randomly selecting 27 group members for groups with more than 27 members. In addition, to further show that results are robust to changes in group size, we randomly selected 15 members from each group. Random selection was conducted 1,000 times, and average ICC(1) values were calculated. In the replication both with 27 and with 15 members, the rank order pattern of item ICC(1) values over time was identical to those in Table 5. Furthermore, estimates themselves were also nearly identical to results from the full analyses. The largest ICC(1) difference across conditions was |0.0044|. Overall, these results suggest (a) temporal rank order stability in item ICC(1) values and (b) evidence that ICC(1) values are robust to unbalanced data and different group sizes.

Example 3: Safety Climate

For our third example, we examined a subset of data described in detail by Hofmann and Mark (2006), which includes the perceptions of nurses and safety outcomes. Our study was more focused and therefore less subject to missing data, so we used data from a larger sample than the original report. Our sample included

⁶ The fact that only 3 of the 6 items were asked on the survey demonstrates the relatively common practice of reducing scale length to minimize respondent burden particularly common in repeated-measures designs.

Table 3
Individual- and Group-Level Psychometric Properties of the Leadership Scale

Item	Abbreviated items descriptions	Total ($\alpha = .93$)		Within ($\alpha = .92$)		Between ($\alpha = .98$)	
		Alpha without	ITC	Alpha without	ITC	Alpha without	ITC
1	Push me to do my best	.92	.72	.91	.68	.98	.92
2	Are motivated	.92	.70	.91	.65	.98	.92
3	Keep me motivated	.91	.81	.90	.79	.97	.94
4	Make me want to do my best	.91	.80	.90	.78	.97	.97
5	Inspire me to try harder	.91	.81	.90	.80	.97	.95
6	Help build my confidence	.92	.77	.90	.76	.98	.92
7	Do not care about how I do*	.92	.68	.91	.64	.98	.89
8	Do not care about training us*	.92	.68	.91	.63	.97	.94
9	Give up on us*	.92	.67	.91	.61	.98	.88

Note. $N = 2,199$ individuals in 47 units. ITC is item-total correlation. Items marked with asterisks are reverse-coded.

1,987 individuals in 117 units of 62 hospitals. We measured unit size and safety outcomes from information provided by 1,313 respondents who reported the frequency with which they had experienced needle sticks. We examined item-level properties from 750 nurses who answered all nine safety climate items.

Measures

To demonstrate the consequences of considering item-level ICC(1) information, we constructed four alternative scales of *safety climate* from the 9 items used by Hofmann and Mark (2006). The items were drawn from a previously used scale (Mueller, DaSilva, Townsend, & Tetrick, 1999) adapted from Zohar (1980). We formed 3-item scales from the pool of items on the basis of (1) traditional psychometric scale development criteria, (2) Between CFA, (3) Between ML-CFA, and (4) item-level ICC(1) values. The items underlying each scale are listed in Tables 6 and 7.

We summed nurses' reports of the number of times in the prior three months they had been stuck by a sharp object or needle, to calculate unit *needle sticks*. This dependent variable differs slightly from that used by Hofmann and Mark (2006); their measure was a count of reported incidents from archival data. Because the outcome variable was captured as a unit's sum, we accounted for unit size as the number of nurses who answered the needle stick question.

Analysis and Results

As with the previous leadership example, we subjected the items to a broad range of individual-level and group-level psychometric tests.⁷ Based on the various approaches, we constructed four 3-item scales to measure safety climate. We then compared the scales' effectiveness in predicting the safety outcome, needle sticks. In practice, shortening the scale to only 3 items runs the risk of raising content validity issues, but the practice is fairly common and we use it simply to demonstrate the implications of different procedures for item selection.

Individual-level psychometric properties. Consistent with common practice, we calculated alpha values for the full, 9-item scale and with each item removed. We also calculated item-total correlations for each item. The most favorable values were items 2, 3, and 7 (see Table 6). The first data column in Table 7 provides

individual-level CFA results ($\chi^2 = 225.22$; $p < .001$; $CFI = 0.92$; $SRMR = 0.08$) and suggests a different set of items (7, 8, and 9).

Multilevel and proposed scale development practices. Standardized factor loadings from several variants of CFA are displayed in Table 7. Results of the CFA Within ($\chi^2 = 196.53$; $p < .001$; $CFI = 0.93$; $SRMR = 0.07$) and ML-CFA analyses ($\chi^2 = 219.56$; $p < .001$; $CFI = 0.93$; $SRMR_{within} = 0.07$; $SRMR_{between} = 0.32$) were consistent with the overall CFA results. Likewise, the CFA-Between results ($\chi^2 = 95.63$; $p < .001$; $CFI = 0.87$; $SRMR = 0.12$) also mirrored the CFA-Within results, reflecting the fact that group-level CFAs inherit effects from lower levels of analysis (see Dyer, Hanges, & Hall, 2005, p. 153; Muthén, 1990).

The ML-CFA-Between shows substantial differences from individual-level analyses in the magnitude of loadings, their variance, and the rank order of the items. The values differ not only from the individual item-total results, but also from CFA-Between values. The 3-item scale suggested by the ML-CFA-Between included items 2, 7, and 8. Two safety items (4 and 5) had negative factor loadings. This is surprising for item 4, which has a nominally high negative coefficient for the factor loading ($\lambda = -.55$; $SE = 1.39$; $p = .70$). The finding, however, is less surprising in light of the large standard error, and the ICC(1) value of .01 that is indistinguishable from zero. Finally, ICC(1) values (last column of Table 7) varied substantially across items and suggested a reduced scale with items 1, 3, and 7.

On the basis of various approaches, we constructed four 3-item scales to measure safety climate (individual-level psychometric with items 2, 3 and 7; CFA-Between with items 7, 8 and 9; ML-CFA-Between with items 2, 7 and 8; and ICC(1)-based with items 1, 3 and 7). The scale based on the CFA-Between had the lowest ICC(1) value (.14), and the others were roughly equal (.17 for psychometric and ICC-based values and .18 for the ML-CFA-Between value). ICC(1) confidence intervals for all scales overlapped substantially. The highest alpha value among the four scales was observed for the Between CFA ($\alpha = .79$), and the

⁷ Consistent with the multidimensional nature of the scale as confirmed by Hofmann and Mark (2006), we conducted our single-level and multi-level confirmatory factor analyses with freed error covariance among the first three items, which constituted the strongest subordinate factor in our data.

Table 4
Multilevel Properties of the Leadership Scale

Item	Abbreviated items descriptions	CFA loadings (Total)	CFA loadings (Within)	CFA loadings (Between)	ML-CFA loadings (Within)	ML-CFA loadings (Between)	ICC(1) [CI]
1	Push me to do my best	.75	.72	.92	.62	.45	.10 [.06, .15]
2	Are motivated	.72	.67	.93	.60	.58	.21 [.14, .28]
3	Keep me motivated	.86	.84	.96	.80	.48	.11 [.06, .16]
4	Make me want to do my best	.86	.85	.98	.80	.46	.08 [.04, .12]
5	Inspire me to try harder	.86	.86	.97	.82	.45	.08 [.04, .12]
6	Help build my confidence	.82	.82	.95	.82	.43	.07 [.04, .10]
7	Do not care about how I do*	.68	.65	.90	.65	.49	.11 [.06, .16]
8	Do not care about training us*	.68	.62	.93	.55	.56	.17 [.10, .23]
9	Give up on us *	.66	.61	.88	.57	.65	.21 [.14, .29]

Note. $N = 2,199$ individuals in 47 units. Items marked with asterisks are reverse-coded. CFA Within is confirmatory factor analysis based on individual deviations from group means ($N = 2,199$). CFA Between is confirmatory factor analysis based on unit average values of each item ($N = 47$). ML-CFA is multilevel confirmatory factor analysis. ICC is intra-class correlation. CI is 95% confidence interval.

lowest was found in the scale based on ML-CFA-Between ($\alpha = .62$). The alpha values for the psychometrically based scale ($\alpha = .76$) and the ICC-based scale ($\alpha = .74$) exceeded the usual cutoff point of .70. We next tested the predictive validity of the various subscales.

Predicting safety outcomes. Consistent with Hofmann and Mark (2006), we predicted needle sticks with a negative binomial regression, accounting for the nesting of units in hospitals by estimating clustered robust standard errors. We conducted separate regression equations for each of the four subscales, predicting needle sticks from standardized group-level scores of safety climate with unit size as a covariate. The reported coefficients are incident rate ratios (IRRs), which can be interpreted as the proportional risk of an outcome in one unit compared with a comparable unit one standard deviation lower on safety climate (Simon, 2001).

For comparison, a linear combination of all 9 items predicted the safety outcome negatively and the scale was significant at a 90% confidence level ($IRR = .59$, $SE = .19$, $p < .10$) implying a 41% decrease in needle sticks. The coefficient for the psychometric scale ($IRR = .57$, $SE = .20$, ns) implied a decrease in accidents but was not statistically significant. The coefficient for the scale based on CFA-Between was also negative and not statistically significant ($IRR = .62$; $SE = .19$; ns). The coefficient for the scale based on ML-CFA-Between was negative and significant at the 90% confidence level ($IRR = .58$; $SE = .17$; $p < .10$). The ICC(1)-based scale was negatively and statistically significant at the 95% level ($IRR = .44$; $SE = .15$; $p < .05$).

The relationships between standardized values of the three scales and reports of needle sticks are displayed in Figure 1. The form of the relationships was not substantially different. Pairwise

comparisons revealed that the effect sizes did not statistically significantly differ from each other, which is not surprising given the overlap in the constructs and items. Although chance may partly account for the finding that only the ICC-based scale met the 95% threshold for significance, the overall pattern involving both the ML-CFA-Between and the ICC-based scales illustrate the value of considering items that detect group differences.

Discussion

Our review of the literature showed that current measurement validation procedures for group-level measures typically focus on reliability and CFA information conducted at the individual level. These analyses provide limited information about whether items differentiate between groups, so using individual-level results to select items runs the risk of creating scales that fail to optimize group differentiation as part of the overall validation process. As a consequence, group-level relationships may frequently be underestimated, lowering our ability to fully appreciate the theoretical and practical value of group-level variables. Our examples show that ICC(1) values from each item in a scale provide valuable and unique information about between-group variability. Item ICC(1) values appear temporally stable and reasonably invariant to differences in group size and can easily be incorporated into the validation process.

One advantage to incorporating information about item-level ICC(1) values into the measurement validation process is that the procedure is computationally simple. ML-CFA-Between provided similar (but not identical) information to that offered by ICC(1), but ML-CFA is computationally intensive, and its predictive validity was nominally (though not substantially) lower than the

Table 5
Multilevel Properties of the Cohesion Scale

Item	Abbreviated items descriptions	Time 1 ICC(1) [CI]	Time 2 ICC(1) [CI]	Time 3 ICC(1) [CI]
1	The members of my platoon are cooperative with each other	.17 [.12, .25]	.20 [.14, .29]	.16 [.11, .24]
2	The members of my platoon know they can depend on each other	.12 [.08, .18]	.17 [.12, .25]	.14 [.09, .21]
3	The members of my platoon stand up for each other	.11 [.07, .17]	.16 [.11, .24]	.13 [.08, .19]

Note. $N = 2,249$ at Time 1; 2,248 at Time 2; 2,274 at Time 3. 47 units at all times. ICC(1) is the intra-class correlation. CI is 95% confidence interval.

Table 6

Individual- and Group-Level Psychometric Properties of the Unit Safety Climate Scale

Item	Abbreviated items descriptions	Total ($\alpha = .79$)		Within ($\alpha = .77$)		Between ($\alpha = .81$)	
		Alpha without	ITC	Alpha without	ITC	Alpha without	ITC
1	Job duties prevent nurses from acting safely	.75	.58	.73	.54	.77	.63
2	Job duties interfere with ability to comply	.75	.62	.72	.58	.76	.66
3	Job duties interfere with levels of safety	.75	.60	.72	.56	.77	.64
4	Best nurses expect other nurses to ensure safety	.80	.23	.77	.25	.82	.16
5	Best nurses care about patient/nurse safety	.79	.26	.77	.25	.82	.20
6	Nurses remind each other to ensure safety	.79	.33	.76	.30	.80	.41
7	Nurse manager insures adequate resources	.75	.59	.73	.53	.76	.69
8	Nurse manager views violations very seriously	.77	.47	.75	.42	.79	.46
9	Nurse manager emphasizes safety practices	.76	.56	.73	.54	.78	.59

Note. $N = 750$ individuals in 117 units. ITC is item-total correlation.

scale based on the ICC(1) in our third example. Furthermore, the estimation of a ML-CFA model requires access to more groups than the estimation of ICC(1) values. For instance, authors such as Hox and Maas (2001) and Preacher, Zyphur, and Zhang (2010) have pointed out that 50 groups represents a small sample for ML-CFA. On the basis of our examples, we anticipate that ML-CFA and ICC(1) results would show similar consistency in other samples. If results differed, researchers would need to consider assumptions underlying the statistical models, noting that when the number of groups is small, ICC(1) results may be more defensible.

Multilevel Measurement Validation

Multilevel measurement validation, like all measurement validation, is complex and multifaceted. The framework outlined by G. Chen et al. (2004) provides a firm foundation for approaching multilevel measurement, but we suggest that the framework can be enhanced by further emphasizing the importance of between-group differentiation. Finding temporal stability in ICC(1) values (Example 2) implies that some items are consistently more effective at differentiating groups than other items. To some degree, this finding is not surprising because a similar phenomenon is observed with respect to individual items used to assess individual constructs. One of the first steps in measurement validation is to

generate a large pool of items with the knowledge that some items will be discarded and others will be retained. In multilevel data, similar procedures can be accomplished using item-level ICC(1) results.

As with best practices in developing individual-level constructs, we encourage the use of multiple samples in multilevel measurement validation. Using multiple samples will help reduce the possibility that differences in item-level ICC(1) values are a product of one specific context. We believe that subtle differences in the wording of items produces differences in ICC(1) values that will be consistent across contexts, but it is possible that something about one specific context (e.g., the military) led to some of the patterns we observed.

We are not proposing that researchers use ICC(1) values alone. Indeed, it is relatively easy to conceive of items that can differentiate groups but have little construct validity. For instance, group members should be quite accurate estimating the age of their immediate supervisor, and because supervisors often differ on age, an ICC(1) of this measure would presumably be high. Nonetheless, including supervisor age in a scale measuring a collective construct such as cohesion would reduce the scale's validity. As another example, teams likely differ in meaningful ways with respect to affect expression (Barsade & O'Neill, 2014); however,

Table 7

Multilevel Properties of the Unit Safety Climate Scale

Item	Abbreviated items descriptions	CFA loadings (Total)	CFA loadings (Within)	CFA loadings (Between)	ML-CFA loadings (Within)	ML-CFA loadings (Between)	ICC(1) [CI]
1	Job duties prevent nurses from acting safely	.35	.28	.43	.28	.83	.11 [.04, .19]
2	Job duties interfere with ability to comply	.38	.31	.43	.31	.90	.09 [.05, .17]
3	Job duties interfere with levels of safety	.35	.28	.41	.27	.87	.10 [.06, .18]
4	Best nurses expect other nurses to ensure safety	.29	.32	.22	.34	-.55	.01 [*, *]
5	Best nurses care about patient/nurse safety	.36	.37	.31	.41	-.09	.05 [.02, .12]
6	Nurses remind each other to ensure safety	.43	.42	.41	.45	.44	.04 [.01, .12]
7	Nurse manager insures adequate resources	.75	.71	.83	.72	.95	.18 [.12, .26]
8	Nurse manager views violations very seriously	.69	.66	.69	.66	.98	.07 [.03, .14]
9	Nurse manager emphasizes safety practices	.79	.78	.85	.78	.89	.09 [.04, .17]

Note. $N = 750$ individuals in 117 units. CFA Within is confirmatory factor analysis based on individual deviations from group means. CFA Between is confirmatory factor analysis based on unit average values of each item ($N = 117$). ML-CFA is multilevel confirmatory factor analysis. ICC is intra-class correlation. CI is 95% confidence interval.

* ICC(1) CI interval estimates for Item 4 returned implausible estimates.

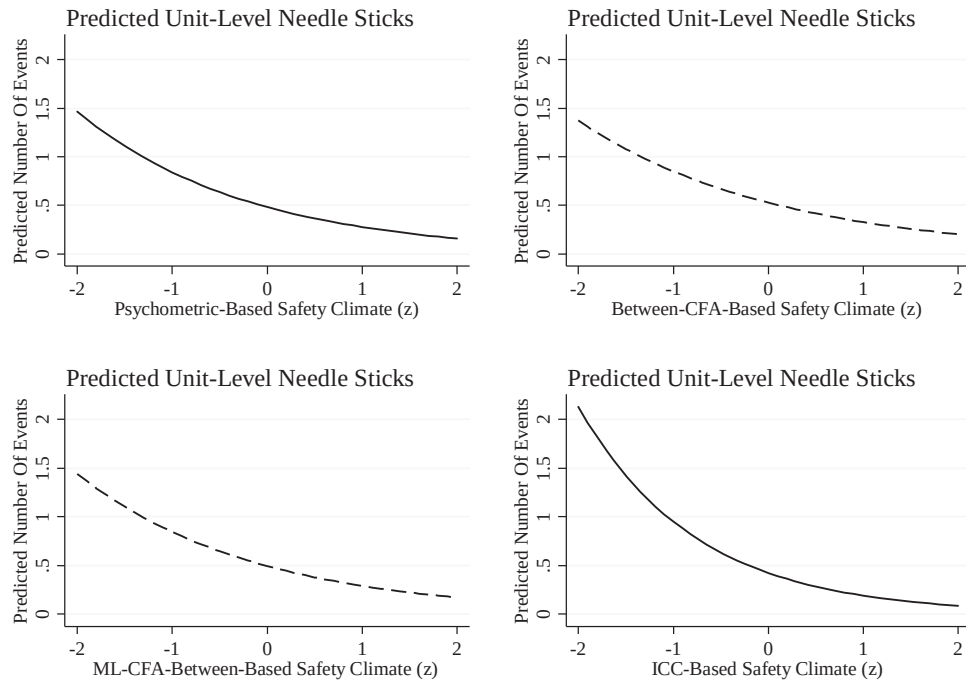


Figure 1. Prediction of unit-level needle sticks from various measures of safety climate.

we would not recommend including an item about group affect if affect was not a core component of the larger construct. Considering item-level ICC(1) values within the existing framework proposed by G. Chen et al. (2004) helps balance trade-offs among the criteria of reliability, validity, content validation, and between-group differentiation.

Ideally, the procedure of examining item-level ICC(1) values would be incorporated in the initial stages of scale development. Despite the fact that we illustrated the effects of examining ICC(1) when reducing scales, we remain agnostic as to whether reducing items in scales is an acceptable practice. We do, however, echo comments made by Aguinis and colleagues, who strongly encourage authors using altered scales (either through changing the referent group or omitting items) to not only describe how and why scales were changed, but also provide additional evidence for the scale's validity (Aguinis, Ramani, & Alabduljader, 2018; Aguinis & Vandenberg, 2014). Our position is simply to emphasize that item-level ICC(1) values provide one additional piece of evidence important for understanding scale validity, and thus item-level ICC(1) values should be examined and reported.

Interestingly, routinely examining and reporting item-level ICC(1) values may also help researchers design group-level measures that are more conceptually and empirically sound. For example, some of the items with the lowest ICC(1) values in the safety climate example focused on "best nurses" (items 4 and 5). Top performers may be key in setting the safety climate; however, group members may have difficulty consistently identifying the top performers. Such inconsistencies would contribute to the items' poor ability to differentiate among groups.

Likewise, Example 1 illustrated the importance of considering item-referent (the unit, the respondent) from a measurement view (see also Arthur et al., 2007; Klein et al., 2001), and we recom-

mend researchers developing group-level scales consider evidence that referent-shift items can improve agreement and are often more construct valid (Wallace et al., 2016). That said, even when the referent is consistent (e.g., cohesion items in Table 5), items can show considerable differences in the ability to detect group differences. Ultimately, the wording of items should match theory and be consistent within a scale.

Limitations

Our examples represent typical multilevel constructs (leadership, cohesion, and safety climate) across two settings. Nonetheless, we do not know if our examples would be mirrored in other constructs used by the field. That said, we believe the cumulative evidence provides a reasonable basis to conclude that ICC(1) differences in items are common, and we have seen similar item-level differentiation in other data such as engagement data from firms. Ultimately though, given the simplicity of our recommendation, there appears little downside to encouraging researchers to routinely consider item-level ICC(1) information.

Finally, it is important to emphasize that the ICC(1) is a "blind metric," in the words of one of our reviewers. In the example involving safety climate, we cannot really conclude that the scale created from the three items with the highest ICC(1) values were the best at capturing "climate." The items could have been the best at capturing the effects of other group-level differences like training programs or the effectiveness of nurse managers (leaders). This emphasizes our earlier point that item-level ICC(1) results need to be interpreted within the larger context of validity to understand what is being assessed to produce an appropriate scale.

Conclusion

In summary, we have provided a simple way to increase the quality of scales designed to assess group-level constructs. Our proposed change of encouraging researchers to examine item-level ICC(1) values as part of the overall validation process has the potential to produce group-level measures that are more sensitive to differences among groups. Optimizing the measurement of group differences should help facilitate detecting group-level relationships and help advance multilevel theory and testing.

References

- Adler, A. B., Bliese, P. D., Pickering, M. A., Hammermeister, J., Williams, J., Harada, C., . . . Ohlson, C. (2015). Mental skills training with basic combat training soldiers: A group-randomized trial. *Journal of Applied Psychology, 100*, 1752–1764. <http://dx.doi.org/10.1037/apl0000021>
- Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals, 12*, 83–110. <http://dx.doi.org/10.5465/annals.2016.0011>
- Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 569–595. <http://dx.doi.org/10.1146/annurev-orgpsych-031413-091231>
- Arthur, W., Jr., Bell, S. T., & Edwards, B. D. (2007). A longitudinal examination of the comparative criterion-related validity of additive and referent-shift consensus operationalizations of team efficacy. *Organizational Research Methods, 10*, 35–58. <http://dx.doi.org/10.1177/1094428106287574>
- Aryee, S., Walumbwa, F. O., Seidu, E. Y. M., & Otake, L. E. (2012). Impact of high-performance work systems on individual- and branch-level performance: Test of a multilevel model of intermediate linkages. *Journal of Applied Psychology, 97*, 287–300. <http://dx.doi.org/10.1037/a0025739>
- Bandura, A. (2000). Exercise of human agency through collective efficacy. *Current Directions in Psychological Science, 9*, 75–78. <http://dx.doi.org/10.1111/1467-8721.00064>
- Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology, 83*, 377–391. <http://dx.doi.org/10.1037/0021-9010.83.3.377>
- Barsade, S. G., & O'Neill, O. A. (2014). What's love got to do with it? A longitudinal study of the culture of companionate love and employee and client outcomes in a long-term care setting. *Administrative Science Quarterly, 59*, 551–598. <http://dx.doi.org/10.1177/0001839214538636>
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.
- Bliese, P. D., Halverson, R. R., & Schriesheim, C. A. (2002). Benchmarking multilevel methods in leadership: The articles, the model, and the data set. *The Leadership Quarterly, 13*, 3–14. [http://dx.doi.org/10.1016/S1048-9843\(01\)00101-1](http://dx.doi.org/10.1016/S1048-9843(01)00101-1)
- Bliese, P. D., Maltarich, M. A., & Hendricks, J. L. (2018). Back to basics with mixed-effects models: Nine take-away points. *Journal of Business and Psychology, 33*, 1–23. <http://dx.doi.org/10.1007/s10869-017-9491-z>
- 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. <http://dx.doi.org/10.1037/0021-9010.83.2.234>
- Chen, G., & Bliese, P. D. (2002). The role of different levels of leadership in predicting self- and collective efficacy: Evidence for discontinuity. *Journal of Applied Psychology, 87*, 549–556. <http://dx.doi.org/10.1037/0021-9010.87.3.549>
- Chen, G., Farh, J. L., Campbell-Bush, E. M., Wu, Z., & Wu, X. (2013). Teams as innovative systems: Multilevel motivational antecedents of innovation in R&D teams. *Journal of Applied Psychology, 98*, 1018–1027. <http://dx.doi.org/10.1037/a0032663>
- Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multi-level construct validation. *Research in Multi-Level Issues, 3*, 273–303. [http://dx.doi.org/10.1016/S1475-9144\(04\)03013-9](http://dx.doi.org/10.1016/S1475-9144(04)03013-9)
- Chen, X. P., Liu, D., & Portnoy, R. (2012). A multilevel investigation of motivational cultural intelligence, organizational diversity climate, and cultural sales: Evidence from U.S. real estate firms. *Journal of Applied Psychology, 97*, 93–106. <http://dx.doi.org/10.1037/a0024697>
- Chen, Z., Zhu, J., & Zhou, M. (2015). How does a servant leader fuel the service fire? A multilevel model of servant leadership, individual self identity, group competition climate, and customer service performance. *Journal of Applied Psychology, 100*, 511–521. <http://dx.doi.org/10.1037/a0038036>
- Chiniara, M., & Bentein, K. (2016). Linking servant leadership to individual performance: Differentiating the mediating role of autonomy, competence and relatedness need satisfaction. *The Leadership Quarterly, 27*, 124–141. <http://dx.doi.org/10.1016/j.leaqua.2015.08.004>
- Cohen, A., Doveh, E., & Nahum-Shani, I. (2009). Testing agreement for multi-item scales with the indices $r_{WG(j)}$ and $AD_{M(j)}$. *Organizational Research Methods, 12*, 148–164. <http://dx.doi.org/10.1177/1094428107300365>
- Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software, 20*, 1–18. <http://dx.doi.org/10.18637/jss.v020.i02>
- Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly, 16*, 149–167. <http://dx.doi.org/10.1016/j.leaqua.2004.09.009>
- Foran, H. M., & Adler, A. B. (2013). Trainee perceptions of drill sergeant qualities during basic combat training. *Military Psychology, 25*, 577–587. <http://dx.doi.org/10.1037/mil0000023>
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72–91. <http://dx.doi.org/10.1037/a0032138>
- Gibson, C. B., Randel, A. E., & Earley, P. C. (2000). Understanding group efficacy: An empirical assessment of multiple assessment methods. *Group & Organization Management, 25*, 67–97. <http://dx.doi.org/10.1177/1059601100251005>
- Hammer, L. B., Kossek, E., Bodner, T., & Crain, T. (2013). Measurement development and validation of the family supportive supervisor behavior short-form (FSSB-SF). *Journal of Occupational Health Psychology, 18*, 285–296. <http://dx.doi.org/10.1037/a0032612>
- Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 89–127). Mahwah, NJ: Erlbaum.
- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*, 967–988. <http://dx.doi.org/10.1177/014920639502100509>
- Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121. <http://dx.doi.org/10.1177/109442819800100106>
- Hirschfeld, R. R., Jordan, M. H., Feild, H. S., Giles, W. F., & Armenakis, A. A. (2006). Becoming team players: Team members' mastery of teamwork knowledge as a predictor of team task proficiency and observed teamwork effectiveness. *Journal of Applied Psychology, 91*, 467–474. <http://dx.doi.org/10.1037/0021-9010.91.2.467>
- Hofmann, D. A., & Mark, B. (2006). An investigation of the relationship between safety climate and medication errors as well as other nurse and

- patient outcomes. *Personnel Psychology*, 59, 847–869. <http://dx.doi.org/10.1111/j.1744-6570.2006.00056.x>
- Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8, 157–174. http://dx.doi.org/10.1207/S15328007SEM0802_1
- Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, 86, 3–16. <http://dx.doi.org/10.1037/0021-9010.86.1.3>
- Kozlowski, S. W. J., & Klein, K. J. (2000). A multi-level approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3–90). San Francisco, CA: Jossey-Bass.
- Lang, J. W. B., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence in groups using longitudinal multilevel methods. *Personnel Psychology*, 71, 255–281. <http://dx.doi.org/10.1111/peps.12260>
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11, 815–852. <http://dx.doi.org/10.1177/1094428106296642>
- Mathieu, J. E., Kukenberger, M. R., D'Innocenzo, L., & Reilly, G. (2015). Modeling reciprocal team cohesion-performance relationships, as impacted by shared leadership and members' competence. *Journal of Applied Psychology*, 100, 713–734. <http://dx.doi.org/10.1037/a0038898>
- Mikolon, S., Kreiner, G. E., & Wieseke, J. (2016). Seeing you seeing me: Stereotypes and the stigma magnification effect. *Journal of Applied Psychology*, 101, 639–656. <http://dx.doi.org/10.1037/apl0000060>
- Mueller, L., DaSilva, N., Townsend, J., & Tetric, L. (1999 April 30–May 2). *An empirical evaluation of competing safety climate measurement models*. Paper presented at the 14th Annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.
- Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology*, 43, 131–143. <http://dx.doi.org/10.1111/j.2044-8317.1990.tb00930.x>
- Nishii, L. H. (2013). The benefits of climate for inclusion for gender-diverse groups. *Academy of Management Journal*, 56, 1754–1774. <http://dx.doi.org/10.5465/amj.2009.0823>
- Podsakoff, P. M., & MacKenzie, S. B. (1994). An examination of the psychometric properties and nomological validity of some revised and reduced substitutes for leadership scales. *Journal of Applied Psychology*, 79, 702–713. <http://dx.doi.org/10.1037/0021-9010.79.5.702>
- Podsakoff, P. M., MacKenzie, S. B., & Ahearne, M. (1997). Moderating effects of goal acceptance on the relationship between group cohesiveness and productivity. *Journal of Applied Psychology*, 82, 974–983. <http://dx.doi.org/10.1037/0021-9010.82.6.374>
- Podsakoff, P. M., MacKenzie, S. B., & Fetter, R. (1993). Substitutes for leadership and the management of professionals. *The Leadership Quarterly*, 4, 1–44. [http://dx.doi.org/10.1016/1048-9843\(93\)90002-B](http://dx.doi.org/10.1016/1048-9843(93)90002-B)
- Podsakoff, P. M., Niehoff, B. P., MacKenzie, S. B., & Williams, M. L. (1993). Do substitutes for leadership really substitute for leadership? An empirical examination of Kerr and Jermier's situational leadership model. *Organizational Behavior and Human Decision Processes*, 54, 1–44. <http://dx.doi.org/10.1006/obhd.1993.1001>
- Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, 49, 518–543. <http://dx.doi.org/10.1080/00273171.2014.933762>
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209–233. <http://dx.doi.org/10.1037/a0020141>
- Priesemuth, M., Schminke, M., Ambrose, M. L., & Folger, R. (2014). Abusive supervision climate: A multiple-mediation model of its impact on group outcomes. *Academy of Management Journal*, 57, 1513–1534. <http://dx.doi.org/10.5465/amj.2011.0237>
- Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, 16, 295–330.
- Schwab, D. P. (1980). Construct validity in organizational behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 2, pp. 3–43). Greenwich, CT: JAI Press.
- Scott, B. A., & Judge, T. A. (2009). The popularity contest at work: Who wins, why, and what do they receive? *Journal of Applied Psychology*, 94, 20–33. <http://dx.doi.org/10.1037/a0012951>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. <http://dx.doi.org/10.1037/0033-2909.86.2.420>
- Simon, S. D. (2001). Understanding the odds ratio and the relative risk. *Journal of Andrology*, 22, 533–536.
- Sliter, K. A. (2013). Development and validation of a measure of workplace climate for healthy weight maintenance. *Journal of Occupational Health Psychology*, 18, 350–362. <http://dx.doi.org/10.1037/a0033132>
- Tasa, K., Taggar, S., & Seijts, G. H. (2007). The development of collective efficacy in teams: A multilevel and longitudinal perspective. *Journal of Applied Psychology*, 92, 17–27. <http://dx.doi.org/10.1037/0021-9010.92.1.17>
- Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel construct measures. *Organizational Research Methods*, 17, 77–106. <http://dx.doi.org/10.1177/1094428113517008>
- Wallace, J. C., Edwards, B. D., Paul, J., Burke, M., Christian, M., & Eissa, G. (2016). Change the referent? A meta-analytic investigation of direct and referent-shift consensus models for organizational climate. *Journal of Management*, 42, 838–861.
- Williams, J., Brown, J. M., Bray, R. M., Anderson Goodell, E. M., Olmstead, K. R., & Adler, A. B. (2016). Unit cohesion, resilience, and mental health of soldiers in basic combat training. *Military Psychology*, 28, 241–250.
- Zohar, D. (1980). Safety climate in industrial organizations: Theoretical and applied implications. *Journal of Applied Psychology*, 65, 96–102. <http://dx.doi.org/10.1037/0021-9010.65.1.96>
- Zweber, Z. M., Henning, R. A., & Magley, V. J. (2016). A practical scale for multi-faceted organizational health climate assessment. *Journal of Occupational Health Psychology*, 21, 250–259. <http://dx.doi.org/10.1037/a0039895>
- Zyphur, M. J., Kaplan, M. F., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, 12, 127–140. <http://dx.doi.org/10.1037/1089-2699.12.2.127>

Received July 5, 2017

Revision received July 18, 2018

Accepted July 19, 2018 ■