Organizational Research Methods OnlineFirst, published on January 18, 2008 as doi:10.1177/1094428107309322

Organizational Research Methods
Volume XX Number X
Month XXXX xx-xx
© 2008 Sage Publications
10.1177/1094428107309322
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Composing Group-Level Constructs From Individual-Level Survey Data

Heleen van Mierlo
Erasmus University Rotterdam and Radboud University Nijmegen
Jeroen K. Vermunt
Tilburg University
Christel G. Rutte
Eindhoven University of Technology

Group-level constructs are often derived from individual-level data. This procedure requires a composition model, specifying how the lower level data can be combined to compose the higher level construct. Two common composition methods are direct consensus composition, where items refer to the individual, and referent-shift consensus composition, where items refer to the group. The use and selection of composition methods is subject to a number of problems, calling for more systematic work on the empirical properties of and distinction between constructs composed by different methods. To facilitate and encourage such work, the authors present a methodological framework for addressing the distinction between and the baseline psychometric quality of composed group constructs, illustrated by an empirical example in the group job-design domain. The framework primarily represents a developmental tool with applications in multilevel theory building and scale construction, but also in meta-analysis or secondary analysis, and more general, the validation of group constructs.

Keywords: *aggregation; composition models; measuring group constructs*

Groups and teams are an everyday phenomenon in organizational life. Employees are gathered in work groups, project teams, consultation groups, management teams, quality circles, social groups, departments, and so on. Such groups are an inexhaustible source of inspiration for organizational researchers. Although the domain of group research is extremely varied, one issue that concerns most, if not all, group researchers is the measurement of group-level phenomena. Group-level phenomena can be measured in a variety of ways. In the organizational sciences, the most common approach is to collect individual survey responses and aggregate those to the group level (Klein, Conn, Smith, & Sorra, 2001; Mossholder & Bedeian, 1983; Rousseau, 1985).

When collecting and aggregating individual data to measure a group phenomenon, researchers implicitly or explicitly adopt a composition model (Klein et al., 2001; Rousseau,

Authors' Note: This study was supported by Fatigue at Work, a research program by the Dutch Organization for Scientific Research (Grant 580-02.206). Christel G. Rutte is now at the Faculty of Social Sciences, Tilburg University, Netherlands. Correspondence concerning this article should be addressed to Heleen van Mierlo, Institute of Psychology, Erasmus University Rotterdam, T12-39, P.O. Box 1738, 3000 DR Rotterdam, Netherlands; e-mail: vanmierlo@fsw.eur.nl.

1985). A composition model describes how a construct that is operationalized at one level of analysis is related to another form of that construct at a different level (James, 1982; Rousseau, 1985). It is thus a formal representation of the method that is used for composing a group-level construct from individual data. A composition model could describe how individual psychological climate is related to organizational climate, how self-efficacy is related to collective efficacy, or how individual task design is related to group task design. Composition models apply to pretty much any situation in which lower level information is used to make inferences about a higher level construct. While our current focus is on individuals in groups, our line of reasoning equally applies to, for example, departments in organizations, pupils in classrooms, subordinates and leaders, or residents in countries.

In 1998, David Chan proposed a typology of five different composition models: additive, direct-consensus, referent-shift consensus, dispersion, and process models. The direct-consensus and referent-shift consensus composition models are most common in organizational research. A direct-consensus model uses “within-group consensus of the lower-level units as the functional relationship to specify how a construct conceptualized and operationalized at the lower level is functionally isomorphic to another form of the construct at the higher level” (Chan, p. 237). The meaning of the higher level construct lies in the consensus among the lower level units. Direct-consensus composition involves two steps. First, the two constructs of interest are defined and operationalized (one at the lower and one at the higher level). As an example, one might use individual responses to a psychological climate measure to operationalize individual psychological climate, and operationalize team climate as the average of the individual responses within a group. The second step consists of specifying how and under what conditions the individual scores may be summarized to represent the higher level construct. A common condition for aggregation of lower level scores is within-group agreement. A minimum amount of agreement is required to demonstrate that averaging individual responses yields a reliable and valid group-level construct. Group members should, for example, provide similar responses to the psychological climate measure to allow the group to be described by its average score.

The referent-shift consensus model is similar to the direct-consensus model in that the group-level construct is based on individual-level responses. The models differ in that referent-shift consensus involves an additional step: Prior to composing the group-level construct, the referent of the individual-level measure is changed. As such, the group-level construct is not composed directly from the individual-level construct but from an altered version. Instead of asking group members to assess individual psychological climate, one would ask them to assess the climate of their group. This altered version of the individual-level construct would then be aggregated to the group level, again after establishing sufficient agreement among group members.

The widespread application of direct-consensus and referent-shift consensus composition methods in the measurement of group phenomena is subject to a number of challenges or problems. First, the most appropriate composition method is not always evident. Although some domains of group research have developed widespread consensus as to the most appropriate composition model (e.g., group efficacy is by definition concerned with the question “can we do this task?” so that referent-shift composition is the obvious choice), in many other instances the meaning of constructs in terms of their level of reference is much more ambiguous. The domain of leadership, for example, involves a host of different theoretical approaches, referring to different levels of analysis and presupposing different composition models, whereas attention for level-of-analysis issues is relatively new to this field (Yammarino, Dionne, & Uk Chun, 2002; Yammarino, Dionne, Uk Chun, & Dansereau, 2005). The same is true for the domain of job design, where the level of theory is often not specified or ambiguous so that selecting appropriate measurement procedures is rather complicated (Van Mierlo, Rutte, Kompier, & Doorewaard, 2005).

Second, many authors do apply consensus composition procedures but fail to explicate composition issues either conceptually, methodologically, or both. Although many domains of group research are making considerable progress in the development of multilevel approaches with explicit attention for composition issues, composition issues are not yet fully integrated into the group research tradition. This leaves us with a large body of research, old and new, that does not, or only partly, addresses composition issues, resulting in conceptual and methodological ambiguity with regard to level issues, and obscuring the interpretation of study results.

Last, data aggregation is sometimes inspired, at least in part, by pragmatic, data-analytical considerations. There are many situations in which the conceptual model is framed at the group level whereas individual data collection is the only feasible data collection procedure. In such situations, the a priori focus on group-level theory and analysis might instigate "automatic" aggregation without due consideration of the individual character of the measurement procedure. As such, data-analytical considerations sometimes outweigh the need for conceptual work on composition issues.

Until now, we know only little about the implications of these problems because empirical work on the distinction between direct-consensus and referent-shift composition is largely lacking. In the case of a pronounced empirical distinction, that is, when direct-consensus composition and referent-shift consensus composition yield distinct, unique constructs, the implications of the above-mentioned problems are clearly much more far reaching compared to a situation where the two composition procedures yield constructs that are very similar. As far as we know, two previous studies addressed this issue, yielding inconsistent results. Schriesheim (1979) concluded that "group oriented and individual oriented leadership descriptions are nearly identical" (p. 353), suggesting no difference between direct-consensus and referent-shift composition. Klein et al. (2001) examined the wording of survey items as antecedent of within-group agreement about group members' work situation and supervision. They found that, in specific occasions, the use of items that refer to the group rather than to the individual situation increased within-group agreement, and tentatively concluded that items with a group referent may be better able to capture group-level constructs. Klein et al. (2001) and Schriesheim (1979) employed different techniques to compare composition methods. In our view, more systematic work is needed to gain insight into the implications of the problems related to the use and selection of composition models. With the purpose of facilitating and encouraging such work, in the following, we present a methodological framework for examining the properties of and empirical distinction between direct-consensus and referent-shift consensus composition. This framework represents a straightforward tool that requires no unfeasible study designs or extensive (multilevel) statistical knowledge. Such a framework is relevant in a number of ways. First and foremost, it can serve as an important developmental aid in the scale-construction phase of group research to complement conceptual work on composition issues. Most notably, the framework helps shed light on issues related to the reliability and construct validity of composed group-level constructs and helps select the most appropriate composition method. Second, the approach is of general interest for empirical work in domains of group research where conceptual developments have not (as yet) indicated a single, most appropriate composition model, either because different composition models may be equally valid, or because conceptual work with regard to composition issues still needs to be developed. Third, our framework can be of use for those undertaking literature reviews, meta-analysis, or secondary analysis. Because previous studies in many domains of group research applied a variety of composition methods, meaningful summary and interpretation of results requires insight into the statistical implications of the use of distinct composition procedures. Fourth and finally, our approach provides an outline for systematic examination of the statistical properties of a single composition procedure.

Our framework comprises five complementary steps that, together, can be used to examine the existence and nature of the empirical distinction between direct-consensus and referent-shift composition. Steps 1 and 2 address whether or not a distinction exists, whereas the remaining three steps examine the reliability and validity of the resulting group-level constructs. We introduce these steps below and subsequently illustrate their application with an example in the domain of group job design. Figure 1 provides an overview of the framework, summarizing key information for each step.

Step 1: Similarity of Constructs: Factor Analysis Between Groups

The purpose of this first step is to examine the extent to which direct-consensus composition and referent-shift consensus composition yield distinct group-level constructs. One method of examining this is factor analysis (cf. Schriesheim, 1979). If factor analysis on the aggregated group-level data would yield a structure in which the direct-consensus and referent-shift items load on distinct factors, this would indicate that direct-consensus and referent-shift composition yield distinct group constructs. The use of factor analysis involves a number of choices with regard to, for example, the factor analytical procedure and factor retention criteria. These choices should be attuned to the specific context and purposes of the analysis (for a detailed discussion see, e.g., Fabrigar, MacCallum, Wegener, & Strahan, 1999; Hayton, Allen, & Scarpello, 2004; Lance, Butts, & Michels, 2006). We briefly discuss our choices in this respect when presenting our empirical illustration of the framework.

Step 2: Similarity of Constructs: Correlations

The potential distinction between two group-level constructs can further be quantified by their correlation. The standard correlation, based on the raw scores, provides a first indication of the extent to which team members differentiate between their own situation and that of their team. The larger this correlation, the larger the overlap between the two measures. However, the standard (raw-scores) correlation can present a distorted image of the actual situation because it does not take into account the clustering in the data structure. Therefore, it will often be more informative to separately examine the correlation

Figure 1
Framework for Addressing the Distinction Between and the Baseline Psychometric Quality of Composed Group Constructs

Step 1: Similarity of Constructs: Factor Analysis Between Groups

Rationale: Direct-consensus and referent-shift items loading on different factors suggests distinct constructs.

Procedure: Between-group exploratory factor analysis (EFA) on direct-consensus and referent-shift items (simultaneously).

Data: Between-group: aggregated group-level data $(\overline{X}_j)$.

Interpretation: Joint inspection of component loadings, Scree plot, and eigenvalues or parallel analysis (Hayton, Allen, & Scarpello, 2004).

Step 2: Similarity of Constructs: Correlations

Rationale: $r_{(\text{individual})}$= Extent to which raw scores reflect a distinction between direct-consensus and referent-shift.

$r_{(\text{within-groups})}$ = Extent to which respondents distinguish between own and group situation.

$r_{(\text{between-groups})}$ = Extent to which direct-consensus and referent-shift items and corresponding scales are distinct after aggregation of individual responses.

Procedure: Correlations between direct-consensus and referent-shift items and scale means.

Data: Individual: Original individual-level data (X_{ij}).

Within groups: Individual deviation scores from group means $(X_{ij} - \overline{X}_j)$.

Between groups: Aggregated group-level data $(\overline{X}_j)$.

Interpretation: Large correlations indicate lack of discriminant validity.

*Step 3: Reliability and Construct Validity: Variance W*ithin and Between Groups

Rationale: Comparison of between-group variance to within-group variance and total variance provides an indication of the primary source of variation in the data and thereby of there liability of the group-level construct.

Procedure:

$$ICC(1) = \frac{MSB - MSW}{MSB + (N-1)MSW};$$

$$ICC(2) = \frac{MSB - MSW}{MSB} \text{ or } \frac{N[ICC(1)]}{1+(N-1)ICC(1)}$$

$$\text{WABA I: } \eta_{BX} = \sqrt{\frac{SS_B}{SS_T}};\ \eta_{WX} = \sqrt{\frac{SS_W}{SS_T}};\ E = \frac{\eta_{BX}}{\eta_{WX}};\ F = E^2\frac{N-J}{J-1}\ (J = N_{\text{groups}})$$

ICC(1), ICC(2), and WABA I analyses can, for example, be based on a one-way ANOVA withgroup as random factor.

Data: Original individual-level data (X_{ij}).

Interpretation: ICC(1): Statistical significance and absolute value (see e.g., Bliese, 2000).

ICC(2): Interpreted as a reliability measure. Cutoff criterion depends on intended use of group scores (see Lance, Butts, & Michels, 2006; Nunnally, 1978).

WABA I: Practical significance based on 30º (group level if $E \geq 1.73$, within-group level if $E < 0.577$) or 15º test (group level if $E \geq 1.30$, within-group level if $E < 0.77$); statistical significance based on *F* test or 1/*F* test (Dansereau, Alutto, & Yammarino, 1984).

(continued)

Figure 1 (Continued)

Step 4: Reliability & Construct Validity: Agreement Within Teams

Rationale: Higher agreement within groups suggests a more valid group construct.

Procedure: $r*_{wg(j)} = 1 - \frac{\bar{s}_x^2}{_{EU}\,s^2}$ (Lindell, Brandt, & Whitney, 1999); $\bar{s}_{xj}^2$ = mean of observed variances on J items;
s_{EU}^2 = Expected variance under uniform distribution = (A2–1)/12 (A = alternatives in response scale).

Data: Original individual-level data (X_{ij}).

Interpretation: For 5-point scales, $-1.00 \le r^*_{wg(J)} \le 1.00$; for 4-point scales, $-.80 \le r^*_{wg(J)} \le 1.00$; $r^*_{wg(J)} = 0$ in case of random response; and $r^*_{wg(J)} = 1.00$ in case of maximum agreement. Cutoff scores for aggregation and statistical significance testing are subject to continuing debate. Dunlap et al. (2003) provide $r^*_{wg(J)}$ significance levels for various combinations of sample size and number of categories.

Step 5: Construct Validity: Factor Analysis Within Groups

Rationale: A clear factor structure on within-group level suggests systematic individual differences.

Procedure: Within-group exploratory factor analysis on direct consensus and on referent shift items (separately).

Data: Within-group: individual deviation scores from group means $(X_{ij} - \bar{X}_j)$.

Interpretation: The clearer the factor structure and the higher the explained variance, the larger the systematic response differences between individual group members. Factor retention decisions might be based on joint inspection of component loadings, Scree plot, and eigenvalues or parallel analysis (Hayton et al., 2004).

Note: ICC = intraclass correlation coefficient; MSB = between-group mean square; MSW = within-group mean square; *N* in the ICC formulas = group size; WABA = within and between analysis; η_{BX} = between-group eta correlation; η_{WX} = within-group eta correlation; SS_B = between-group sum of squares; SS_W = within-group sum of squares; *E* = E-test of practical significance; *F* = *F* test of statistical significance; *N* in WABA formula for *F* test = individual-level sample size; *J* in WABA formula for *F* test = group-level sample size (the number of groups).

within groups and *between groups* (cf. at the aggregated group level; e.g., Robinson, 1950; Snijders & Bosker, 1999; Yammarino, 1990). Within groups, the correlation between direct-consensus and referent-shift items expresses the extent to which respondents differentiate between their individual perception of their own individual situation and that of their group. Between groups, this correlation indicates the extent to which direct-consensus and referent-shift items and their corresponding scales are distinct after aggregation of the individual responses to the group level. Even if the measures are distinct at the individual level, once aggregated they may be very similar. Concluding, the extent to which direct-consensus and referent-shift items and their corresponding scales represent distinct group-level constructs is reflected in their correlation, both overall, within groups, and between groups (see Figure 1).

There are no clear-cut decision rules for interpreting the resulting correlation coefficients. Probably the best-known standard for interpreting correlation coefficients is provided by J. Cohen (1988): .10 for a small correlation, .30 for a medium correlation, and

.50 for a large correlation. In addition, according to Kenny (1998), a correlation of .85 or larger would indicate poor discriminant validity. Based on these figures, we would tend to consider a correlation between .50 and .70 as large but still indicating considerable distinctness, a correlation between .70 and .85 as indicating substantial overlap, and a correlation of .85 or higher as indicating a definite lack of discriminant validity. This proposition is, however, by no means intended as a golden standard. Researchers with specific theoretical interest in the nature of the distinction and overlap between constructs composed by direct-consensus and referent-shift consensus might consider further exploration of, for example, differential relationships and predictive validity of the two constructs.

Step 3: Reliability: Variance Within and Between Groups

Steps 1 and 2 address the distinction between direct-consensus and referent-shift group constructs but not the nature of this distinction. As such, they provide no information on the reliability and validity of the group constructs that are composed by the two methods. As such, the purpose of Step 3 is to examine the extent to which direct-consensus and referent-shift consensus composition yield reliable group constructs. If all groups have similar scores on a group measure, the measure does not differentiate between groups and thus is not a reliable group-level construct (e.g., Klein & Kozlowski, 2000). The intraclass correlation coefficient (ICC) provides an indication of the proportion of group-level variance. There are numerous versions of ICC (see Shrout & Fleis, 1979), two of which are of particular interest in group research: ICC(1) and ICC(2). ICC(1) equals the correlation between the values of two randomly drawn individuals from a single randomly drawn group. This correlation is commonly interpreted as the proportion of variance in a target variable that is accounted for by group membership (Bliese, 2000; McGraw & Wong, 1996; Snijders & Bosker, 1999). ICC(2) represents the reliability of the group mean scores and varies as a function of ICC(1) and group size, so that large group sizes can result in high ICC(2) values, even if ICC(1) values are low (Bliese, 2000). ICC(1) and ICC(2) values are often used to assess whether aggregation to the group level is appropriate. Statistically significant ICC(1) values suggest dependence in the data structure, indicating that individual-level analyses would be inappropriate, whereas high ICC(2) values indicate reliable between-group differences, supporting aggregation to the group level. For a group-level construct to be reliable, it should yield significant ICC(1) and acceptable ICC(2) values. As such, the higher ICC(1) and ICC(2), the larger the extent to which the construct is shared by group members and the more reliable the resulting group construct.

A complementary method of assessing the primary source of variance is presented by the within- and between-analysis approach (WABA; Dansereau, Alutto, & Yammarino, 1984; Dansereau, Cho, & Yammarino, 2006; Dansereau & Yammarino, 2006). Of primary interest here is the WABA I procedure that assesses, for each variable separately, whether it varies primarily between groups, within groups, or both. WABA I involves estimating between and within eta-correlations (see Figure 1 for formulas), tested for practical and statistical significance with, respectively, E- and *F* tests. As explained above, a reliable group-level construct should be relatively homogeneous within groups and heterogeneous between groups, as indicated by higher between-eta correlations relative to within-eta correlations.

Step 4: Construct Validity: Agreement Among Group Members

ICC expresses the consistency in responses of members of the same group compared to members of different groups and is therefore generally considered a reliability coefficient. It does not represent the extent to which group members provide exactly identical ratings of the subject of interest, typically referred to as (within-group) interrater agreement (IRA). If we are really measuring a group-level construct, we would expect group members to provide similar ratings in an absolute sense, at least more similar than under conditions of random response (e.g., Bliese, 2000; A. Cohen, Doveh, & Eick, 2001; Klein et al., 2001). A common measure of IRA is the $\mathrm{r}_{wg(J)}$index for multiple items (James, Demaree, & Wolf, 1984). This index is obtained by comparing the observed variance in a group on a set of items to the variance that would be expected if group members would respond randomly. The mathematical properties of $\mathrm{r}_{wg(J)}$are subject to considerable debate. To address some of these problems, Lindell, Brandt, and Whitney (1999) proposed an alternative version named $\mathrm{r}^{*}_{wg(J)}$. An important advantage of $\mathrm{r}^{*}_{wg(J)}$ is that, unlike $\mathrm{r}_{wg(J)}$, rating scales with large numbers of items do not result in overestimation of true agreement (Lindell et al., 1999). We therefore recommend the use of $\mathrm{r}^{*}_{wg(J)}$ over that of $\mathrm{r}_{wg(J)}$. Interpretation of $\mathrm{r}^{*}_{wg(J)}$ values is somewhat problematic. The cutoff of .70 that is used by many authors as a rule of thumb for acceptable agreement is arbitrary and not grounded in theory (Lance et al., 2006) and the chi-square test of statistical significance is not robust for the small sample sizes that characterize a lot of research on groups and teams (Lindell et al., 1999). Dunlap, Burke, and Smith-Crowe (2003) presented an alternative statistical significance test, based on Monte Carlo procedures and presented an overview of $\mathrm{r}^{*}_{wg(J)}$ significance levels for various combinations of sample size and number of categories. Based on the current state of the literature on assessing IRA, we would tentatively propose the use of Lindell et al.'s (1999) $\mathrm{r}^{*}_{wg(J)}$, interpreted based on the absolute $\mathrm{r}^{*}_{wg(J)}$ value in concurrence with a test of statistical significance based on Dunlap et al. (2003). We should note here once again that we do not mean to present a golden standard. There are other options, and the literature on IRA indices for assessing within-group agreement is rapidly developing.

Step 5: Construct Validity: Factor Analysis Within Groups

The extent to which a group construct indeed captures a "true" group phenomenon represents yet another indication of the validity of an aggregated group-level construct. This is, however, rather challenging to establish because researchers primarily resort to aggregate measures when direct measurement at the group level is unfeasible. Because collecting objective group-level data is often difficult, time-consuming, and expensive, assessing the relationship between an aggregated group-level construct and its "true" group-level counterpart is often impossible.

As an alternative, factor analysis may provide an indication of the extent to which a group-level construct captures a group-level phenomenon (e.g., Dansereau & Yammarino, 2006). In direct-consensus composition, the group-level construct is based on a measure of the individual situation. Individual measures are, by definition, designed to capture individual differences. A valid individual measure (e.g., individual autonomy or self-efficacy) should differentiate between members of different groups, but preferably also between members of the same group. Although members of the same group are typically more similar in many respects than members of different groups, their individual situations and perceptions still differ. An individual measure is designed to be sensitive to these differences, even though such individual differences (within and between groups) might at times be somewhat mitigated by group or contextual influences (e.g., recruitment strategies, leadership, or organizational culture). Factor analysis is an elementary way to examine response structures. Because individual measures are originally designed to detect differences between group members that result from "true" differences in their individual situations, factor analysis within groups should yield a clear one-factor structure (provided one used a valid one-dimensional measure).

In referent-shift consensus composition, on the other hand, the items are not meant to measure individual differences between group members. On the contrary, because they all assess the same group phenomenon, all members of a work group will be expected to answer such items in a similar way. If we assume that researchers who employ referent-shift composition are interested in a "true" or "objective" group-level phenomenon and that they trust group members to be able to make a reliable judgment on this aspect of their group, differences between members of the same work group represent measurement error. Such differences may arise, for example, if group members are not all equally well informed of the group phenomenon and are thus somewhat imprecise in their assessment. A statistical assumption that underlies most analysis techniques is the assumption of nonsystematic measurement error. If all variance in the individual responses within a group is indeed attributable to nonsystematic error, factor analysis within groups should yield no meaningful structure.

Empirical Example

Together, these five steps constitute a framework that may provide more insight into some methodological properties of and empirical differences between group constructs as composed by direct-consensus and referent-shift consensus composition. To illustrate the framework, we now present an empirical example that stems from job design theory. In the domain of job design, aggregation of individual responses is widely applied (Campion, Papper, & Medsker, 1996; Edwards, Scully, & Brtek, 2000; Hackman & Oldham, 1980), whereas thorough consideration of multilevel issues often seems to be lacking (Van Mierlo et al., 2005). In our example, we concentrate on the job design dimensions of autonomy and variety. Autonomy is the extent to which an employee can control work processes, whereas variety is the extent to which a job requires different skills and talents (Hackman & Oldham, 1975; Karasek, 1990). Both constructs were developed to assess individual job characteristics and were originally framed at the individual level. Later, they were applied to the level of the work group. Autonomy and variety are considered important determinants of team effectiveness and represent primary criteria for the design of self-managing teams (e.g., Campion et al., 1996; Cordery, 1996;

Langfred, 2005; Van Mierlo et al., 2005). It is interesting to note that apart from the shift in level, application of autonomy and variety to the group level typically has not resulted in changes in the meaning or definition of the constructs. As such, *group-level autonomy* refers to the extent to which a work group can control work processes and variety to the extent to which to the group task requires different skills and talents. The literature includes direct-consensus and referent-shift consensus approaches of group-level autonomy and variety. In case of a direct-consensus approach, the average level of individual autonomy and/or variety is (implicitly or explicitly) considered an adequate representation of the autonomy or variety of the group as a whole. This could be an appropriate approach in a context where within-team homogeneity is a valid assumption. As an example, in the domain of self-managing teamwork "employee discretion over decisions such as task assignments, methods for carrying out the work and scheduling of activities" is often considered a key feature of such teams. This suggests a high average level of individual autonomy in such teams, while the level of theory in this field is unmistakably that of the work team (e.g., Parker, Wall, & Cordery, 2001; Van Mierlo et al., 2005). A referent-shift consensus approach to group autonomy and variety would be appropriate in the more general case where group members are considered reliable informants about the group as a whole.

As such, conceptually, both composition approaches to group autonomy and variety are justifiable, although such explicit justification is rarely provided in the literature on group job design.

Method

Sample

Our example is based on data from a large data set about job characteristics in a group context and various aspects of psychological well-being. These data were collected in five team-based health care organizations in the Netherlands. Two were domiciliary care organizations, while the remaining three were nursing homes. The work groups in these organizations were referred to as "self-managing teams" and had well-defined group tasks, usually consisting of the care for all clients in a specific area or ward. Group members met regularly and often received training in, for example, work planning systems or communication skills. Survey data were available from 753 members from 80 work groups, representing an average response rate of 63%. Two groups were excluded because of a large number of missing values, and two others because only one member responded to the questionnaire. The remaining sample consisted of 733 members of 76 work groups. Response rates per team varied from 30% to 100% ($M = 70\%$), and the number of respondents per team ranged from 2 to 22, with a mean of 9.64 ($SD = 5.12$). Teams did not differ significantly with respect to average age or tenure. Most respondents were female (93%), and the average age was 41 years ($SD = 10.62$). Examination of ICC(1) values for organizational membership yielded no substantial between-organization differences, with ICC(1) values of .01 for direct-consensus and for referent-shift variety, .02 for direct-consensus autonomy, and .05 for referent-shift autonomy. Based on these results, we pooled our data across the five organizations.

Measures

All measures were taken from the Dutch Questionnaire on the Experience and Evaluation of Work (VBBA), a self-administered survey instrument developed to assess individual perceptions of the work situation. Previous research has established the psychometric quality of this instrument (Van Veldhoven et al., 2002). All items were answered on a 4-point response scale, ranging from 0 (*never*) to 3 (*all the time*). A complete overview of the survey items is provided in the Appendix.[1]

Individual autonomy was assessed with 11 items asking respondents to indicate the extent to which they could control their work situation, for example, “Can you influence your work pace?” ($\alpha = .86$). Individual variety was assessed with six items, asking respondents to indicate the extent to which their work required the use of different skills and talents, for example, “Is your work varied?” ($\alpha = .77$). These two individual scales represent the direct-consensus model for autonomy and variety.

For the referent-shift consensus model, the individual items were adapted to refer to the work group instead of to the individual employee, for example, “Can your team influence its work pace?” and “Is the work of your team varied?”. Alpha was .89 for referent-shift autonomy and .77 for referent-shift variety.

All respondents received the same survey. As such, each respondent answered the group and individual items. This method may be sensitive to response bias because respondents could be inclined to compare their individual tasks to the group task, instead of making an independent judgment on both. Respondents might, for example, enlarge or attenuate incongruities between their own task and that of their work group. Effort was made to prevent the occurrence of such response bias. Group and individual items were separated by unrelated questions and were never printed on the same page in the survey.

Analyses

The steps constituting our framework encompass multiple levels of data analysis. Step 1 concerns the between-group level. Between-group data can be obtained by averaging the original individual responses to the direct-consensus and the referent-shift items at the level of the work group. Sample size at this level was 76.

Step 5 concerns the within-group level. Within-group data can be obtained by controlling for differences between groups by centering individual scores on their respective group means. Within-group scores thus represent the deviation of individual group members from their work group mean score. The covariance matrix that is calculated from these within-group data is called the “pooled within-group matrix.” This matrix is “pooled” because information from the different groups that was first isolated by calculating deviation scores per group is subsequently combined into a single covariance matrix. This pooled covariance matrix, cleared from all variance between groups, may serve as input for many statistical analyses, including factor analysis (MuthÕn, 1994). Sample size at the within-group level was 733. Steps 3 and 4 involve the original data set, composed of the individual raw scores, with a sample size of 733. Finally, Step 2 involves a combination of all three data levels.

Results

Step 1: Similarity of Constructs: Factor Analysis Between Groups

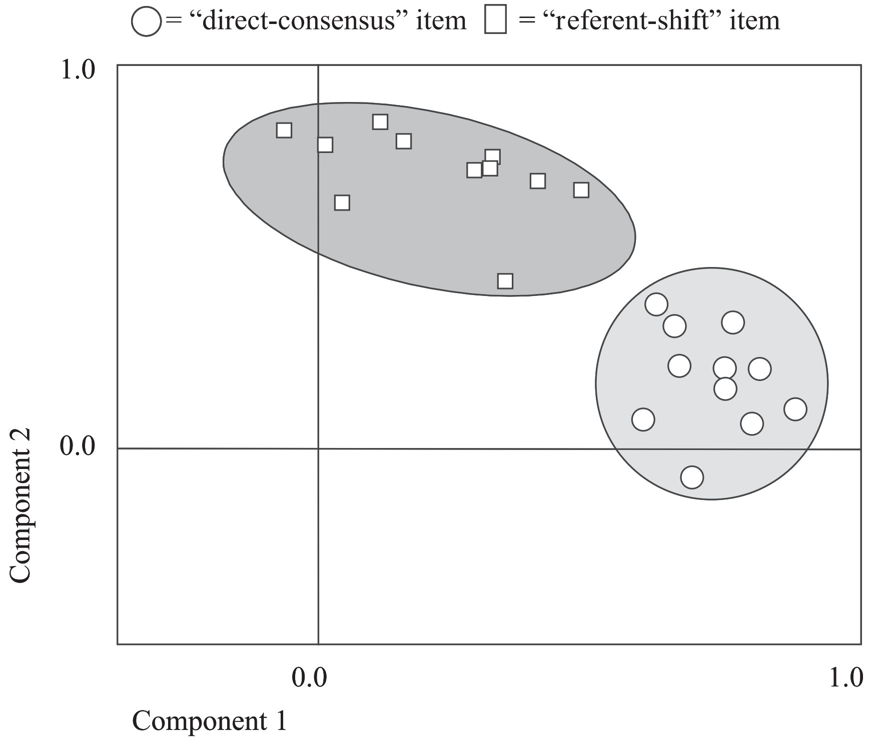
If factor analysis on the between-group scores yields a structure in which the direct-consensus and referent-shift items systematically load on different components, this provides an indication that they represent distinct group constructs. As mentioned above, the use of factor analysis involves a number of important choices. In this empirical example, our choices resulted in the use of exploratory principal components analysis (PCA) with oblique rotation. We took an exploratory rather than confirmatory approach in light of the absence of a priori hypotheses about an expected factor structure (see Hurley et al., 1997). The distinction between PCA and common factor analysis is a fine line that has been subject to discussion (see, e.g., Fabrigar et al., 1999; Velicer & Jackson, 1990). The current context may be interpreted as presenting arguments in favor of both approaches. We employed PCA because our primary interest at this point is to explore the observed data structure, rather than identifying the underlying latent constructs (cf. Fabrigar et al., 1999). Oblique rotation is used in light of the conceptual overlap between direct-consensus and referent-shift consensus constructs that will most likely cause underlying factors to be related. Finally, perhaps the most important decision concerns the selection of factor retention criteria. The common and often default Kaiser's rule (eigenvalue > 1) has consistently been demonstrated to overestimate the number of factors (Hayton et al., 2004; Lance et al., 2006). Therefore, our interpretation of PCA results is based explicitly on joint inspection of the component loadings, Scree plots, and eigenvalues.

With regard to variety, PCA yielded two components that met the eigenvalue > 1 criterion (7.11 and 1.03 and explained variance of 59.32 and 8.60, respectively). The large gap between those eigenvalues in combination with the Scree plot clearly indicated that a one-component model best described the data. As such, the best PCA solution was a one-component model in which all items, direct-consensus and referent-shift, loaded on a single component with an eigenvalue of 7.12 and a percentage of explained variance of 59.32. Component loadings were all above .60. This indicates that the between-group level PCA on the variety items did not distinguish between direct-consensus and referent-shift items. With regard to autonomy, the best PCA solution was a two-component model (eigenvalues of 9.68 and 3.43 and explained variance of 44.00% and 14.59%, respectively). After applying oblique rotation, all direct-consensus autonomy items loaded on the first component, while all referent-shift items loaded on the second. As such, the group-level PCA did distinguish between the direct consensus and the referent-shift consensus model for autonomy, as is clearly illustrated by a schematic representation of the rotated component plot in Figure 2.

Step 2: Similarity of Constructs: Correlations

To examine Step 2, we examined the correlations between the direct-consensus and the referent-shift measures for autonomy and variety at the individual, within-group, and at the between-group level. Table 1 displays the correlations between the individual items and between the scale scores.

Figure 2
Schematic Rotated Component Plot for Autonomy



Note: PCA = principle components analysis. Method = PCA with oblique rotation.

With regard to autonomy, at each level, correlations between the items were remarkably small, especially in light of the identical wording of direct-consensus and referent-shift items. Concentrating on the scale means for autonomy, the percentage of shared variance between direct-consensus and referent-shift autonomy is 24% at the individual level ($r = .49$, $p < .01$), 17% at the within-group level ($r = .41$, $p < .01$), and 31% at the between-group level ($r = .56$, $p < .01$). Individual group members differentiate in their perception of their own autonomy and that of their group (within-group level); and, at the group level, the measures still represent distinct constructs. We therefore conclude that direct-consensus autonomy and referent-shift autonomy represent clearly distinct constructs.

With regard to variety, correlations between the direct-consensus and referent-shift items and scale means were consistently higher than those for autonomy. Still, individuals do seem to differentiate between their own variety and that of their group, as expressed by the moderate individual ($r = .65$, $R^2 = 42\%$, $p < .01$) and within-group ($r = .49$, $R^2 = 24\%$, $p < .01$) correlations. At the aggregated, between-group level, however, the correlations between direct-consensus and referent-shift variety suggest considerable overlap. This is especially true for the correlation of .85 ($p < .01$) between the scale means, representing a

Table 1
Correlations Between Items and Scales Within and Between Groups

	Individual		Within-Group		Between-Group	
Items	Autonomy	Variety	Autonomy	Variety	Autonomy	Variety
Individual 1 x Team 1	.27 **	.50 **	.27 **	.47 **	.29 *	.67 **
Individual 2 x Team 2	.29 **	.57 **	.28 **	.51 **	.35 **	.79 **
Individual 3 x Team 3	.30 **	.52 **	.31 **	.46 **	.32 **	.71 **
Individual 4 x Team 4	.36 **	.39 **	.35 **	.34 **	.55 **	.62 **
Individual 5 x Team 5	.44 **	.39 **	.37 **	.37 **	.71 **	.48 **
Individual 6 x Team 6	.30 **	.48 **	.28 **	.44 **	.42 **	.67 **
Individual 7 x Team 7	.32 **		.32 **		.38 **	
Individual 8 x Team 8	.26 **		.25 **		.32 **	
Individual 9 x Team 9	.15 **		.12 **		.37 **	
Individual 10 x Team 10	.22 **		.21 **		.32 **	
Individual 11 x Team 11	.32 **		.30 **		.47 **	
Scale average	.49 **	.65 **	.41 **	.49 **	.56 **	.85 **

Note: $N_{\text{within}} = 733$, $N_{\text{between}} = 76$.
$^{*}p < .05$. $^{**}p < .01$.

shared variance of 72%. If we were to take into account our measurement error, this percentage would increase even further, implying practically perfect overlap between the two measures. Thus, though individuals did seem to differentiate in their perceptions of their own variety and that of their group (within-group level), at the aggregated group level, direct-consensus and referent-shift variety no longer represent distinct constructs. These findings confirm the results of the PCA.

Step 3: Construct Validity: Variance Within and Between Groups

All results for Step 3 (ICC and WABA I) are displayed in Table 2. With regard to variety, ICC(1) and ICC(2) were considerably higher for the referent-shift measure than for the direct consensus measure: .24 and .67, $F(75, 657) = 3.73$, $p < .001$, versus .14 and .54, $F(75, 657) = 2.50$, $p < .001$. With regard to autonomy, differences between the referent-shift and direct-consensus measure in terms of ICC(1) and ICC(2) were only marginal: .13 and .53, $F(75, 657) = 2.43$, $p < .01$, versus .15 and .55, $F(75, 657) = 2.60$, $p < .001$. WABA I results were not entirely in line with these ICC findings. With regard to variety, the between-eta correlations for referent-shift and direct-consensus variety were practically identical (.48 and .47, respectively) though, with regard to autonomy, the between-eta correlation for the referent-shift measure was slightly larger than that for the direct-consensus measure (.55 and .47, respectively).

Interpretation of the absolute values of the estimates in Table 2 yields an interesting result. On one hand, ICC(1) values are all significant and lie within the range of ICC(1) values commonly encountered in applied field research (e.g., Bliese, 2000), suggesting considerable group-level variance in direct-consensus and referent-shift variety and

Table 2
ICC(1), ICC(2), and WABA I-Results

	Intraclass Correlations			WABA I			
	ICC(1)	ICC(2)	*F* test[a]	η_{between}	η_{within}	E-Test	1/*F* Test[b]
Direct-consensus variety	.14	.54	2.60 **	.48	.88	.55[d]	.38
Referent-shift variety	.24	.67	2.42 **	.47	.89	.53[d]	.41
Direct-consensus autonomy	.15	.55	2.50 **	.47	.88	.53[d]	.40
Referent-shift autonomy	.13	.53	3.73 **	.55	.84	.65[c]	.27

Note: ICC = interclass correlation coefficient; WABA = within and between analysis.
a. *df*(within) = 657; *df*(between) = 75.
b. A parts condition, in which $\eta w > \eta b$, requires an inverse *F* test (1/*F*) with *df* = N – J and J – 1 for the numerator and denominator, respectively (for details, see Dansereau, Alutto, & Yammarino, 1984).
c. E-test for parts significant at 15°.
d. E-test for parts significant at 30°.
$**p < .01$.

autonomy. On the other hand, ICC(2)-values do not reach satisfactory levels for any of the four measures. ICC(2) is generally interpreted as a reliability coefficient. Appropriate cutoff scores for reliability depend on the intended use of the construct but will generally be .80 or higher (see Lance et al., 2006; Nunnally, 1978). As such, in our example, neither composition method yields a group construct that allows reliable comparison of team scores. This ambiguity in ICC results is confirmed by WABA I outcomes. The E-test of practical significance suggests a parts condition for all four measures ($\eta_w > \eta_b$), while the *F* tests of statistical significance indicate that the difference between the within-eta and between-eta correlations is not significant for any of the four constructs. These WABA results imply an equivocal condition at the group level for all four measures and indicate that individual-level effects only may be more likely.

Altogether, though the results for Step 3 do suggest some clustering in the data, variance seems to reside primarily at the individual level.

Step 4: Construct Validity: Agreement Among Group Members

Step 4 concerns $r^*_{wg(J)}$ values for the direct-consensus and referent-shift measures of autonomy and variety. With regard to variety, average $r^*_{wg(J)}$ values over all teams were .68 for the direct-consensus measure and .73 for the referent-shift measure. Seventy-one percent of the groups showed a higher $r^*_{wg(J)}$ value for referent-shift consensus variety compared to direct-consensus variety; however, for most groups, the difference (in either direction) was only small, ranging from .01 to .27 (mean difference = .05). We examined statistical significance only for groups larger than four, because for smaller groups, significance statistics for 4-point response scales are practically meaningless (Dunlap et al., 2003). (With regard to the $r^*_{wg(J)}$ index, statistical significance indicates that the observed agreement is larger than would be expected under the condition of random—uniform—response). Of the 63 groups larger than four, 79% showed significant $r^*_{wg(J)}$ values for direct-consensus variety. For referent-shift consensus variety this was 86%.

Table 3
Principal Components Analysis Within Groups

	Direct-Consensus Autonomy[b]	Referent-Shift Autonomy[b]	Direct-Consensus variety	Referent-Shift Variety[b]
% Variance	39.78	46.32	44.08	43.35
Component loadings				
Item 1	.59	.70	.51	.27
Item 2	.67	.67	.64	.67
Item 3	.58	.54	.79	.80
Item 4	.68	.75	.61	.67
Item 5	.53	.67	.62	.64
Item 6	.71	.74	.78	.76
Item 7	.67	.71		
Item 8	.64	.68		
Item 9	.46	.56		
Item 10	.74	.74		
Item 11	.62	.69		

a. Method: Principal components.

b. Using the common eigenvalue over 1 rule of thumb, we originally obtained two-component solutions for direct consensus autonomy and for referent-shift autonomy and variety. However, because the Scree plots showed large drops only between 0 and 1 components and the eigenvalues were only slightly larger than 1, in all cases a one-component solution was superior. The results of these one-component solutions are displayed in the table.

With regard to autonomy, average $r^*_{wg(J)}$ values over all teams were .66 for the direct-consensus measure and .71 for the referent-shift measure. Seventy percent of the groups yielded a higher $r^*_{wg(J)}$ value for referent-shift consensus variety compared to direct-consensus variety; however, again the difference was small, ranging from .01 to .28 (mean difference = .05). Of the 63 groups larger than four, 73% yielded significant $r^*_{wg(J)}$ values for direct-consensus autonomy. For referent-shift consensus autonomy this was 78%.

Both for autonomy and variety, these results suggest slightly higher within-group agreement for referent-shift consensus composition compared to direct consensus composition; however, for both variables the difference is small. It is interesting to note that all reported values would typically be interpreted as justifying aggregation of individual scores to the group level.

Step 5: Construct Validity: Factor Analysis Within Groups

For the same reasons outlined above under Step 2, we employed exploratory PCA. To examine the final step, we performed PCA on the within-group data. Results are displayed in Table 3. These results show that direct-consensus items for variety and autonomy produced a clear component structure. It is interesting to note that the referent-shift items for variety and autonomy also produced clear component structures that were, in fact, very similar to those for the direct-consensus items. As a side note, the first task variety item was a reverse-coded item, which might explain its lower loading.

These results indicate that within-group component structures for the referent-shift items do not merely represent independent measurement error, but systematic differences. Because variance in within-group data can by definition not be explained by group-level factors, the underlying dimension in our PCA solution for referent-shift autonomy and variety is necessarily related to individual differences.

Conclusion and Discussion

In this article, we presented and illustrated an elementary framework for examining the properties of and the empirical distinction between direct-consensus composition, where survey items refer to the individual situation, and referent-shift consensus composition, where items refer to the work group. This framework is not intended to be exhaustive. Many additional techniques may be used to compare the two types of group constructs. As such, the framework offers a relatively simple tool: a first and basic step in addressing an underexposed and complicated issue. Applications lay, for example, in scale construction in the domain of group research, in meta-analytical or review studies, secondary analysis, and, more generally, in domains of group research where conceptual developments have not (as yet) indicated a single, most appropriate, composition model. In addition, the framework provides an outline for systematic examination of the statistical properties of a single composition procedure.

The framework consists of five complementary steps, addressing two focal concerns: the extent to which direct-consensus composition and referent-shift consensus composition yield distinct group constructs (Steps 1 and 2) and the reliability and construct validity of the resulting group-level constructs (Steps 3, 4, and 5). Next, we review a number of potential outcome scenarios for those two focal concerns and discuss their implications, using the empirical job-design example as an illustration.

Similarity of Constructs

The outcomes of Steps 1 and 2 can be captured in three potential scenarios: (a) no distinction between or within groups, (b) a distinction within-groups but not between-groups, or (c) a distinction within and between groups. In scenario (a), direct-consensus and referent-shift consensus basically measure the same thing. Group members do not distinguish between their own, individual situation and that of the group, and, once aggregated, the constructs overlap to a large extent. In this scenario, statistically, it will make little difference which composition method is used. After all, results of a study employing direct-consensus composition to measure group variety would most likely closely resemble those of a study employing referent-shift consensus composition. In this scenario, either the individual and group-level constructs really are identical, or they are distinct; however, individual group members are not in the position to identify this distinction. The emergence of scenario (a) calls for extended conceptual work on the construct of interest, directed, for example, at the definition and measurement of the group-level construct and the reliance on individual group members to rate the group-level phenomenon.

In scenario (b), the direct-consensus and referent-shift consensus constructs are distinct at the within-group level. That is, individual group members do differentiate between their own situation and that of their team. However, this distinction does not transfer to the between-group level: Once aggregated, the constructs are very similar. This scenario might occur, for example, if group members base their assessment of the group-level construct on an estimated weighted average of the individual situations of all group members. In our empirical example, this might explain the results for the task variety measure, for which scenario (b) emerged. The members of interdependent groups are likely to be familiar with each other's task content and hence to be able to make well-informed inferences of each other's task variety. In this scenario, as in scenario (a), statistically, the selected composition method will make little difference and is unlikely to affect the results of, for example, literature reviews or meta-analyses. Conceptually, however, this scenario does suggest added value for reference-shift consensus composition compared to direct consensus, the methods are not fully interchangeable.

Finally, in scenario (c), direct-consensus and referent-shift consensus composition yield constructs that are clearly distinct. Individuals differentiate between their own situation and that of their group, and this distinction transfers to the group level. In this scenario, direct-consensus and referent-shift composition are not interchangeable under any condition, statistically nor conceptually. The emergence of scenario (c) warrants due caution when interpreting or reanalyzing the results of previous studies and has clear implications for the design and reporting of future studies. Our empirical example suggests that scenario (c) applies to task autonomy. This is an important observation for researchers in the domain of group job design, considering that as yet, it appears not to be a common practice for researchers in the field of group task design to explicate their composition methods (Van Mierlo at al., 2005).

Psychometric Quality of Group-Level Constructs

As indicated by these potential scenarios, the observed distinction between direct-consensus and referent-shift consensus composition, or lack thereof, provides guidance for further conceptual development and is relevant for various research activities relying on previous work. However, it provides no information about the psychometric quality of the resulting group-level constructs. Such information is crucial for any researcher interested in group-level analyses based on an aggregated group construct, whether composed by direct-consensus or referent-shift consensus composition. Therefore, Steps 3, 4, and 5 of our framework address the reliability and validity of group-level constructs. These steps can also be used to compare the reliability and validity of the constructs composed by direct-consensus and referent-shift composition. Potential outcome scenarios with regard to Steps 3, 4, and 5 are (a) good psychometric quality, (b) ambiguous psychometric quality, and (c) insufficient psychometric quality.

Scenario (a) represents the situation we would aim for when interested in group-level measurement and analysis and emerges when Steps 3, 4, and 5 all yield consistent support for reliability and validity. This implies considerable and significant ICC(1) values, indicating that group membership accounts for a meaningful and significant proportion of variance in the group-level construct, large ICC(2) values, indicating reliable group mean scores, significant E- and F tests for WABA I, indicating statistically and practically significant between-group variance relative to within-group variance, significant and large $r^{*}_{WG(j)}$ values for all groups, indicating consistent high levels of absolute agreement among group members and, finally, the absence of a clear factor structure within groups, indicating that within-group variance merely represents random measurement error. This scenario suggests that the composition method yielded a reliable group-level construct that is shared by group members and would, by current standards, be interpreted as supportive of aggregation and subsequent group-level analyses. Still, conceptual debate and development should not end there because, in addition to these baseline psychometric indices, thorough multilevel theory building requires more advanced psychometric study, for example, on discriminant, convergent, and predictive validity of the resulting group construct.

Scenario (b) represents a situation in which Steps 3, 4, and 5 yield mixed results. This scenario calls for a decision rule setting the baseline criteria for adequate psychometric quality. Such decision rules, however, have the habit of being arbitrary and open to debate. In addition, the adequacy of the psychometric quality of a composed group-level construct depends primarily on the intended application of the construct, as is the case in single-level research. Therefore, rather than proposing a decision rule or absolute limit for psychometric quality, we recommend case-by-case consideration of the psychometric properties of aggregated group constructs, based on the combination of indices proposed in Step 3, 4, and 5 of our framework and the general guidelines for the interpretation of those indices. Such case-by-case approach requires explicit and thorough attention for and justification of choices that are made. Among other things, authors should develop and report a priori decision criteria based on the aims of their specific study and pay explicit attention to potential implications of choices made. Scenario (b) is illustrated by the results of our empirical job-design example, where the psychometric results were ambiguous for all four aggregated group constructs (direct-consensus and referent-shift consensus variety and autonomy). ICC and WABA I results for Step 3 suggest an "equivocal" condition for all four aggregated constructs, indicating definite individual-level variance and some degree of group-level variance (Dansereau et al., 1984). In addition, results for Step 4 seem to suggest moderate within-group agreement for the majority, but not for all groups, and, finally, Step 5 suggests systematic within-group differences for all measures. When proceeding with the aggregated construct, caution would clearly be warranted. As an alternative, groups with low within-group agreement could be removed from the sample to increase construct validity. The resulting reduction of within-group variance might also result in more satisfying results for Steps 3 and 5. Of course, the decreased sample size would be a major disadvantage of this approach. In addition, the failure of part of the groups to meet the baseline psychometric criteria might be indicative of conceptual problems with respect to the measurement procedure. Therefore, one may want to consider alternatives for aggregation, such as multilevel modeling (Bryk & Raudenbush, 1992; Snijders & Bosker, 1999), WABA II (Dansereau et al., 1984), or latent variable multilevel modeling for the prediction of group-level outcomes from individual-level data (Croon & Van Veldhoven, 2007).

Finally, scenario (c) represents a situation in which the reliability and construct validity of the composed group-level construct are insufficient. This scenario is implied when the results for Steps 3, 4, and 5 fail to meet the adopted decision criteria in an ambiguous scenario. In this scenario, group-level analyses based on the composed construct are inadvisable and reconsideration of its definition, measurement, and analysis is called for.

Apart from assessing the baseline psychometric quality of composed constructs, our framework allows comparison of the psychometric properties of constructs composed through direct-consensus composition and referent-shift consensus composition. Such knowledge may help select the most appropriate composition method for specific constructs. It is interesting to note that results of our empirical example show no convincing difference in the psychometric properties of direct-consensus autonomy and variety and referent-shift autonomy and variety. Results for Steps 3 and 4 provide some indication of such difference for variety; however, the differences are small, and the emerging pattern is ambiguous for both composition methods. In addition, the results for Step 5 suggest that individual perception differences play an important role in the assessment of direct-consensus and referent-shift consensus composition.

In conclusion, different scenarios may emerge from the application of our framework, and each scenario has its specific implications, for the interpretation of and reliance on the results of previous studies and for the subsequent use of the composed group-level construct(s). An interesting issue that remains concerns the implications of the results for the empirical group job-design example. How should one proceed given the obtained results? Most notably, the ambiguous psychometric results raise some concern regarding the operationalization of group autonomy and variety in terms of either direct-consensus or referent-shift consensus composition. As already mentioned, to address this concern one might, for example, delete from the sample those groups that fail to meet baseline psychometric criteria of reliability and validity and continue analysis with the reduced sample. Alternatively, one may decide to refrain from aggregating the individual scores and instead reframe the autonomy and variety constructs as “individual perception variables.” Note that this strategy would require multilevel analysis because of the clustered data structure. In such a case, common, individual-level analysis can produce distorted estimates (e.g., Snijders & Bosker, 1999). Also, one could employ latent variable multilevel modeling for the prediction of group-level outcomes from individual-level data (Croon & Van Veldhoven, 2007). This procedure explicitly models the measurement model, thus taking into account the unreliability in the measurement procedure. Ultimately, however, in our view, the results of the empirical example expose the need for additional conceptual and developmental work on the measurement of group autonomy and variety, directed at, for example, exploring alternative procedures such as observation, group consensus, or supervisor judgments of job design features.

Having reviewed the various possible scenarios and some implications for our example of group job design, we should note of course that the framework presented in this article is not without its limitations.

In the first place, our aim to keep the framework as accessible as possible has obvious disadvantages. We purposely omitted advanced multilevel techniques and nuances in statistical techniques. One of those nuances concerns the interpretation of the results of PCA. Our interpretation of PCA-results for Steps 1 and 5 is based on joint inspection of the component loadings, Scree plots, and eigenvalues. Although the use of multiple methods helps avoid the risk of overestimating the number of components that typically results from exclusive reliance on Kaiser's rule (eigenvalue > 1) for factor retention (e.g., Lance et al., 2006), parallel analysis presents a still more accurate procedure for factor retention. The interested reader is referred to Hayton et al. (2004) who provided an excellent tutorial on this technique. A second statistical nuance concerns the assessment of within-group agreement. The use of agreement indices has been, and still is, subject to considerable debate. Although a detailed overview of this debate is beyond the scope of the present contribution, we did address some of the disadvantages attached to the original $r_{wg(j)}$ index proposed by James et al. (1984) by proposing the use of $r^*_{WG(j)}$ (Lindell et al., 1999) in combination with a test of statistical significance as proposed by Dunlap et al. (2003). Most notably, $r^*_{WG(j)}$ is not sensitive to the number of items in the scale and is therefore less likely than James et al.'s $r_{wg(j)}$ to overestimate true agreement. The statistical significance test provides important additional information to the absolute $r^*_{WG(j)}$ values, but is somewhat dependent on sample size, and should therefore be interpreted with caution in case of a large number of small groups or large variation in group sizes. In conclusion, as indicated, alternative, at times more advanced, procedures for calculating within-group agreement are available and the debate on agreement indices is still ongoing.

In the second place, it is important to note that group size may affect the difference between the two measurement procedures that we discussed. Members of larger groups might, for example, be more heterogeneous in their assessment of group characteristics than members of smaller groups because they will typically be less aware of the work of many fellow group members. Likewise, group characteristics other than size may affect the properties of and differences between direct-consensus and referent-shift composition measures. Think, for example, of group member physical dispersion, group cohesion, and group development.

In the third place, our framework only represents a general baseline procedure for examining the statistical properties of and distinction between direct-consensus and referent-shift consensus composition. It is intended to encourage and facilitate future efforts in this area. By no means does it provide a comprehensive and self-sufficient approach to the conceptual development of group-level measurement based on composition methods. In addition, our focus on group-level measurement based on individual survey data, and more specifically on direct-consensus and referent-shift consensus composition covers only a limited part of the large arena of group-level measurement. Although these methods still are most common in most domains of group research, the literature shows an increasing interest in alternatives approaches, most notably in group consensus ratings (e.g., Gibson, Randel, & Earley, 2000; Kirkman & Rosen, 1999; Kirkman, Tesluk, & Rosen, 2001; Quigley, Tekleab, & Tesluk, 2007).

In the fourth place, our framework only addresses basic psychometric issues. Further development and assessment of the resulting group-level constructs clearly requires more advanced psychometric study. One important area for further exploration is represented by the nomological net of the group-level constructs. To what extent do constructs composed by direct-consensus and referent-shift consensus demonstrate differential, unique relationships with other variables, such as potential antecedents, correlates, or dependent variables? Some work in this area has already been undertaken, for example, in the area of leadership behavior (Schriesheim, 1979; Yammarino, 1990) and employee perceptions of the work environment (Klein et al., 2001).

In the fifth and final place, it is extremely important to note that the choice for either composition method should always first and foremost be a conceptual one, based on a thorough definition of the constructs of interest (Chan, 1998; Klein, Dansereau, & Hall, 1994). Our current focus on the statistical properties of composition methods should be seen as complementary to such conceptual work.

These limitations notwithstanding, we hope that, ultimately, this article may contribute to increasing transparency into the challenging area of group-level measurement by initiating further empirical research on composition models and by encouraging researchers to carefully consider and report on their group measurement procedures.

Appendix
Overview of Survey Items

The direct-consensus items for autonomy and variety were taken directly from the VBBA, a validated Dutch survey instrument that translates into "Questionnaire on the Experience and Assessment of Work" (Van Veldhoven, De Jonge, Broersen, Kompier, & Meijman, 2002). These original items were adapted to obtain the referent-shift measures. All items were answered on 4-point response scales, ranging from 0 (*never*) to 3 (*all the time*). Please note that all psychometric properties reported on these scales in this article exclusively apply to the Dutch version.

The 11 items used to measure direct-consensus autonomy:

1. Do you have freedom in carrying out your work activities?
2. Do you have influence on the planning of your work activities?
3. Do you have influence on the pace of work?
4. Can you yourself decide how to carry out your work?
5. Can you interrupt your work for a short time if you find it necessary to do so?
6. Can you decide the order in which you carry out your work?
7. Can you participate in the decision about when something must be completed?
8. Can you personally decide how much time you need for a specific activity?
9. Do you yourself resolve problems arising in your work?
10. Can you organize your work yourself?
11. Can you yourself decide on the content of your work activities?

The six items used to measure direct-consensus variety:

1. Do you repeatedly have to do the same things in your work? (RC)
2. Do your work activities require creativity?
3. Are your work activities varied?
4. Does your work require personal input?
5. Do your work activities sufficiently require all your skills and capacities?
6. Do you have enough variety in your work?

The 11 items used to measure referent-shift autonomy:

1. Does your team have freedom in carrying out work activities?
2. Does your team have influence in the planning of the work activities?
3. Does your team have influence on the pace of work?
4. Can your team itself decide how to carry out the work?
5. Can your team interrupt the work for a short time if you find it necessary to do so?
6. Can your team itself decide the order in which to carry out the work?
7. Can your team participate in the decision about when something must be completed?
8. Can your team itself decide how much time is needed for a specific activity?
9. Does your team itself resolve problems arising in its work?
10. Can your team itself organize its work?
11. Can your team itself decide on the content of the work activities?

The six items used to measure referent-shift variety:

1. Does your team repeatedly have to do the same things in its work? (RC)
2. Do the work activities of your team require creativity?
3. Are the work activities of your team varied?
4. Does the work of your team require personal input of the team members?
5. Do the work activities of your team sufficiently require all the skills and capacities of the team members?
6. Does your team have enough variety in its work?

Note: RC = reverse–coded.

References

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

Bryk, A. S., & Raudenbush, S. W. (1982). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Campion, M. A., Papper, E. M., & Medsker, G. J. (1996). Relations between work team characteristics and effectiveness: A replication and extension. *Personnel Psychology*, *49*, 429-451.

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

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the rwg(J) index of agreement. *Psychological Methods*, *6*, 297-310.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cordery, J. L. (1996). Autonomous work groups and quality circles. In M. West (Ed.), *Handbook of workgroup psychology* (pp. 225-246). Chichester, UK: Wiley.

Croon, M. A., & Van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45-57.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach*. Englewood Cliffs, NJ: Prentice Hall.

Dansereau, F., Cho, J., & Yammarino, F. J. (2006). Avoiding the "fallacy of the wrong level": A within and between analysis (WABA) approach. *Group & Organization Management*, *31*, 536-577.

Dansereau, F., & Yammarino, F. J. (2006). Is more discussion about levels of analysis really necessary? When is such discussion sufficient? *Leadership Quarterly*, 17, 537-552.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for RWG and average deviation interrater agreement indexes. *Journal of Applied Psychology*, *88*, 356-362.

Edwards, J. R., Scully, J. A., & Brtek, M. D. (2000). The nature and outcomes of work: A replication and extension of interdisciplinary work-design research. *Journal of Applied Psychology*, *85*, 860-868.

Fabriger, L. R., MacCallum, R. C., Wegener, D. T., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Gibson, C. B., Randel, A. E., & Earley, P. C. (2000). Group efficacy: An empirical test of multiple assessment methods. *Group & Organization Management*, *25*(1), 67-97.

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, *60*, 159-170.

Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Reading, MA: Addison-Wesley.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191-205.

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., et al. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, *18*, 667-683.

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

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

Karasek, R. (1990). Lower health risk with increased job control among white collar workers. *Journal of Organizational Behaviour*, *11*, 171-185.

Kenny, D. A. (1998). *Website*. Available at http://users.rcn.com/dakenny/mfactor.htm.

Kirkman, B. L., & Rosen, B. (1999). Beyond self-management: Antecedents and consequences of team empowerment. *Academy of Management Journal*, *52*, 58-74.

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

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.

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, *19*, 195-229.

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

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. What did they really say? *Organizational Research Methods*, *9*, 202-220.

Langfred, C. W. (2005). Autonomy and performance in teams: The multilevel moderating effect of task interdependence. *Journal of Management*, *31*, 513-529.

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

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

Mossholder, K. W., & Bedeian, A. G. (1983). Cross-level inference and organizational research: Perspectives on interpretation and application. *Academy of Management Review*, *8*, 547-558.

MuthÕn, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376-398.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Parker, S. K., Wall, T. D., & Cordery, J. L. (2001). Future work design research and practice: Towards an elaborated model of work design. *Journal of Occupational and Organizational Psychology*, *74*, 413-440.

Quigley, N. R., Tekleab, A. G., & Tesluk, P. E. (2007). Comparing consensus- and aggregation-based methods of measuring team-level variables: The role of relationship conflict and conflict management processes. *Organizational Research Methods*.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351-357.

Rousseau, D. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. In L. Cummings & B. Saw (Eds.), *Research in organizational behavior* (vol. 7, pp. 1-37). Greenwich, CT: JAI.

Schriesheim, C. A. (1979). The similarity of individual directed and group directed leader behavior descriptions. *Academy of Management Journal*, *22*, 345-355.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Van Mierlo, H., Rutte, C. G., Kompier, M. A. J., & Doorewaard, J. A. C. M. (2005). Self-managing teamwork and psychological well-being: Review of a multilevel research domain. *Group & Organization Management*, *30*, 211-235.

Van Veldhoven, M., De Jonge, J., Broersen, S., Kompier, M., & Meijman, T. (2002). Specific relationships between psychosocial job conditions and job-related stress: A three-level analytic approach. *Work & Stress*, *16*, 207-228.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1-28.

Yammarino, F. J. (1990). Individual- and group-directed leader behavior descriptions. *Educational and Psychological Measurement*, *50*, 739-759.

Yammarino, F. J., Dionne, S., & Uk Chun, J. (2002). Transformational and charismatic leadership: A levels-of-analysis review of theory, measurement, data analysis, and inferences. *Leadership*, 23-63.

Yammarino, F. J., Dionne, S., Uk Chun, J., & Dansereau, F. (2005). Leadership and levels of analysis: A state-of-the-science review. *Leadership Quarterly, 16,* 879-919.

Heleen van Mierlo is an assistant professor of IO-Psychology at the Erasmus University Rotterdam, in the Netherlands. She received her PhD from the Eindhoven University of Technology in collaboration with the Radboud University of Nijmegen, both in the Netherlands. Her current research interests include motivation and wellbeing in groups and teams and multilevel research methods.

Jeroen Vermunt is a professor in the Department of Methodology and Statistics at Tilburg University, the Netherlands. He holds a PhD in social sciences from Tilburg University. He has published extensively on categorical data techniques, methods for the analysis of longitudinal and event history data, latent class and finite mixture models, and latent trait models.

Christel Rutte is a professor in the Faculty of Social and Behavioural Sciences, Department of Psychology and Society at Tilburg University, the Netherlands. Her research interests include group and team processes and time management.