



UvA-DARE (Digital Academic Repository)

Data Aggregation in Team Research

Theoretical Considerations and Practical Recommendations

Nijstad, B.A.; Homan, A.C.; Heerdink, M.W.; van Kleef, G.A.

DOI

[10.1177/20413866251333058](https://doi.org/10.1177/20413866251333058)

Publication date

2025

Document Version

Final published version

Published in

Organizational Psychology Review

License

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

Citation for published version (APA):

Nijstad, B. A., Homan, A. C., Heerdink, M. W., & van Kleef, G. A. (2025). Data Aggregation in Team Research: Theoretical Considerations and Practical Recommendations. *Organizational Psychology Review*, 15(2), 256-287. <https://doi.org/10.1177/20413866251333058>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Data Aggregation in Team Research: Theoretical Considerations and Practical Recommendations

Organizational Psychology Review

2025, Vol. 15(2) 256–287

© The Author(s) 2025


Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/20413866251333058

journals.sagepub.com/home/opr**Bernard A. Nijstad** 

University of Groningen, The Netherlands

Astrid C. Homan 

University of Amsterdam, The Netherlands

Marc W. Heerdink 

University of Amsterdam, The Netherlands

Gerben A. van Kleef

University of Amsterdam, The Netherlands

Abstract

Studying teams comes with notable theoretical and analytical challenges. One major challenge is data aggregation, the process of bringing a variable that was measured at the individual level to the team level of analysis. We provide an accessible resource about data aggregation that integrates theoretical and empirical insights, dispels common misconceptions, and provides practical guidelines for researchers. We stress that decisions about aggregation should be based on theoretical considerations about the nature of variables and their interrelations. Based on the logic of latent variable models, we distinguish reflective and formative team constructs. For reflective team constructs, we discuss issues of within-team agreement and the notion of the conceptual mean. For formative team constructs, we distinguish between well-defined and ill-defined constructs, and offer guidelines for their aggregation. This discussion results in twelve concrete guidelines that can be used by researchers, reviewers, and editors in the design and evaluation of team studies.

Plain Language Summary Title: Creating team properties out of individual data: Guidelines for researchers.

Paper received 5 April 2024. Received revised March 14, 2025. Accepted March 17, 2025

Corresponding author:

Bernard A. Nijstad, Department of HRM/OB, University of Groningen, PO Box 800, 9700AV Groningen, The Netherlands.

Email: b.a.nijstad@rug.nl

Plain Language Summary Title: Plain Language Summary: Researchers who study teams, often want to measure a property of the team. Often, such a property (for example: team cohesion) cannot be directly observed, and researchers therefore ask team members to provide information about this (for example through survey questions). In those cases, researchers have to create one team variable out of the information provided by multiple team members, and this is called data aggregation. This is a very common process in team research, but it is also tricky. In this paper, we make an important distinction between reflective team constructs and formative team constructs. Reflective team constructs are properties of a team for which it is assumed that team members can experience them and that they experience them all in the same way (for example: team cohesion). Formative team constructs do not assume this, but are constructed by the researcher as a summary of individual properties. An example is age diversity, which is directly computed by the researcher based on the ages of the different members. The assumptions for these two types of constructs are very different, and this has implications for how researchers should deal with them. This paper provides a discussion of this and guidelines for researchers about what they should do.

Keywords

Groups/Teams, statistics/methods, data aggregation, diversity & relational demography

Data aggregation in team research: an integration and practical framework

Team dynamics and performance are among the most widely studied topics within the applied psychology, organizational behavior, and management literatures (for reviews: S. G. Cohen & Bailey, 1997; Ilgen et al., 2005; Levine & Moreland, 1990; Mathieu et al., 2008; Mathieu et al., 2017). According to a Web of Science search (March 3, 2025) of the period 2021–2025, in the applied psychology and management literatures alone the word ‘team(s)’ featured in the title of 2092 articles. Despite this popularity among authors and journals, conducting good team research can be challenging. Most of the challenges are rooted in the fact that teams are multilevel systems, in which individuals are nested within teams, and teams are often nested within departments or organizations (Ilgen, 1999; Kenny et al., 2002; Klein et al., 1994; Klein & Kozlowski, 2000; Nijstad, 2009; Rousseau, 1985). The multilevel nature of team studies requires researchers to address the issue of level of analysis in theory, study design, and data analysis.

An important issue in team research is data aggregation, which is the process of bringing a variable that was measured at the lower (individual) level to the higher (team) level of analysis before data analysis. Data aggregation is not only very common in team research, it is also critical: It is fundamentally about the theoretical and empirical relation between lower-level observations and (the meaning of) higher-level constructs (Chan, 1998). As such, data aggregation is not just an analytical affair, but a theoretical one. Unfortunately, it is not uncommon to observe errors in the conceptualization and operationalization of team-level variables and consequently in the analysis and interpretation of associated data. Inadequate conceptualization of variables causes errors that range from suboptimal data aggregation to inappropriate analyses to the unnecessary discarding of data. Such errors threaten the validity of the conclusions drawn from team research and impede insight in psychological processes. We believe many such errors result from misconceptions about data aggregation that are perpetuated due to the lack of an accessible resource that integrates and critically

evaluates key theoretical insights about aggregation, offers best practices that can help researchers make informed decisions, and allows researchers to make the most of their team-level data.

Seminal contributions have been written about the conceptualization of team-level constructs, composition models, measurement models, and multilevel analysis (e.g., Barrick et al., 1998; Chan, 1998; Chen et al., 2004; Kenny et al., 2002; Klein et al., 1994; Klein & Kozlowski, 2000; LeBreton & Senter, 2008; Snijders & Bosker, 1999), and some of these also addressed aspects relevant to data aggregation (most notably Klein & Kozlowski, 2000) and transparent reporting of team-level data (LeBreton et al., 2023). However, there is a need for an updated discussion of data aggregation issues for three main reasons. First, previous discussions (logically) did not include more recent advances around the theorizing about aggregated constructs (e.g., Emich et al., 2021; Waller et al., 2016), their measurement (e.g., Bliese et al., 2018), aggregation statistics (e.g., Biemann & Heidemeier, 2012; LeBreton & Senter, 2008), treatment of missing data (e.g., Allen et al., 2007; Hirschfeld et al., 2013), and the use of latent variables to represent team constructs (Croon & Van Veldhoven, 2007; Lüdtke et al., 2008). Second, the insights that have been developed in different strands of research have not yet been integrated into a single, accessible resource that presents relevant knowledge in a coherent manner. The lack of an integrative and shared understanding of data aggregation in team research has impeded scientific progress by leaving researchers vulnerable to errors that could easily be prevented. Third, and related, there are persistent misconceptions in the field about proper ways of dealing with team-level data that often unnecessarily constrain researchers' leeway in fully exploiting the richness of their data, and limit their ability to gain a deep understanding of team phenomena.

To address these limitations, we provide an integrated approach to data aggregation, inspired by a theoretical distinction that is made in latent variable models: the difference

between reflective and formative latent variable models (e.g., Borsboom et al., 2003; also see Croon & Van Veldhoven, 2007; Lüdtke et al., 2008). We use this distinction to provide a systematic treatment of approaches to data aggregation and integrate this with more basic insights to formulate concrete, stepwise guidelines about data aggregation in team research. We present a theoretically driven and practical framework that can help researchers select proper ways of conceptualizing and operationalizing individual- and team-level variables, decide whether and how to aggregate data, select and interpret appropriate aggregation indices, and redirect statistical analyses based on a joint consideration of a priori theorizing and emerging data patterns. Although some of our guidelines will be familiar to experienced team researchers, we hope that our insights will be valuable for researchers, reviewers, and editors who have less experience with team research and who wish to know more about this important topic.

To illustrate various issues, we also provide a "snapshot" of current practices in data aggregation. To this end, we have systematically examined all papers with the word "team*" in the title and that appeared in five top-tier journals (Academy of Management Journal, Journal of Applied Psychology, Journal of Organizational Behavior, Journal of Management, and Personnel Psychology) in the period of 2016 through 2020. These journals were chosen because they fall within the applied psychology, organizational behavior, or management literatures, have high impact, and publish regularly about team research. In total, this amounted to 104 papers that reported primary, quantitative data on team phenomena. For each study, we established whether data was aggregated, and if so, it was determined per aggregated variable (1) what type of variable it was (e.g., team process, emergent state; see Marks et al., 2001), (2) how it was measured, (3) how it was aggregated, and (4) the aggregation statistics that were reported (when applicable). The coding scheme, the full list of coded papers, and the datafile are here:

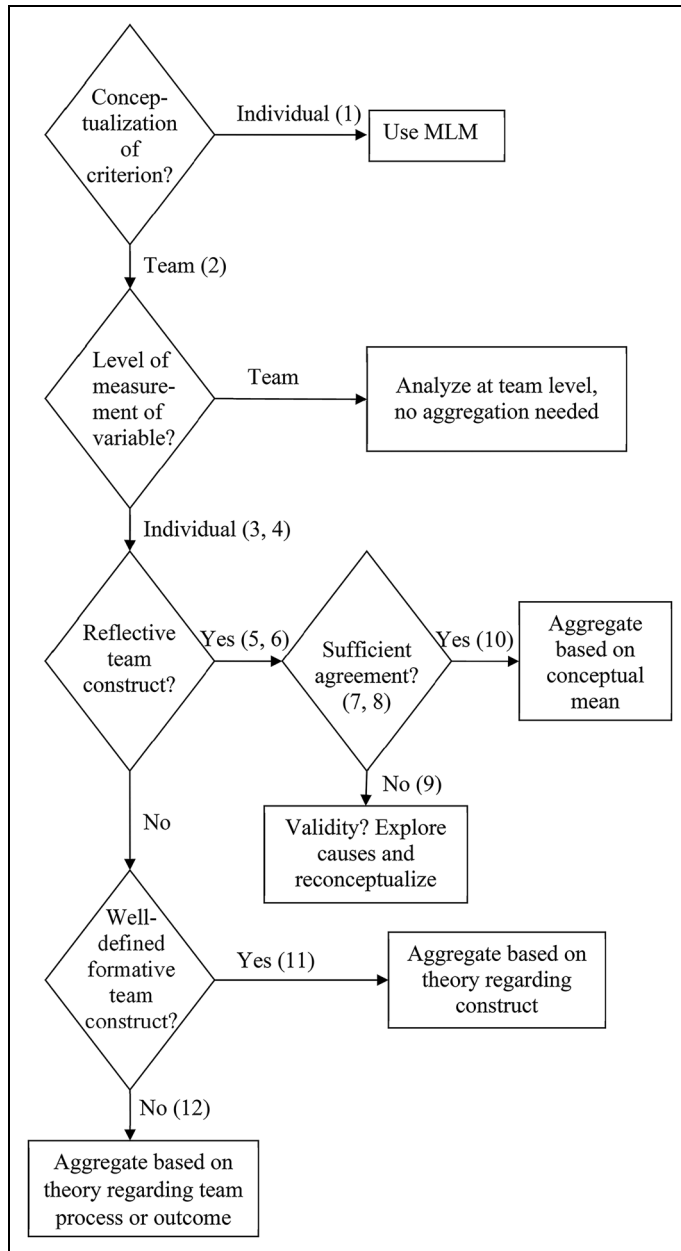


Figure 1. A flow diagram for data aggregation (numbers in parentheses refer to associated guidelines).

https://osf.io/34grp/?view_only=16a0b3cc3d5342a1a519052d30eb2966.

This paper is structured as follows. We first discuss what data aggregation is and when it is appropriate. We then introduce the

conceptual distinction between reflective team constructs (variables for which individual-level scores reflect a team-level construct) and formative team constructs (variables for which a team-level score is formed by a specific

combination of individual-level scores), and discuss the theoretical assumptions underlying them. Subsequently, we more closely consider reflective team constructs, the role of within-team agreement in aggregating these constructs, and the notion of the conceptual mean as an estimate of the reflective team construct. We next discuss formative team constructs, and focus on the best way to aggregate these variables. The result of our discussion is a step-by-step practical guideline for researchers in the form of a flow diagram (see Figure 1).

What is data aggregation and when is it appropriate?

What is data aggregation?

Data aggregation is the process of bringing a variable that was measured at the lower level to the higher level of analysis before data analysis. Data aggregation in team research implies the estimation of a team-level construct based on (multiple) individual-level measurements. This includes (but is not limited to) procedures such as calculating the team-level average, minimum or maximum score, or some measure of (within-team) dispersion or diversity. Data aggregation is very common in team research, and of the 104 papers that we analyzed, in 81% at least one of the variables of the research model (control variables not considered) was aggregated from the individual to the team level. Often, multiple variables were aggregated per study, yielding 301 aggregated variables in 110 different studies.

When a variable is measured at the individual (rather than the team) level, its variance usually reflects a combination of individual-level variance, team-level variance, and measurement error. For example, if we measure individual-level satisfaction with team performance, there can be variance among members within the same team: Some members are more satisfied than others. There can also be variance among teams: In some teams, the average satisfaction is higher than in other

teams (e.g., depending on team performance). In this example, when satisfaction is aggregated to the team level, individual-level variance is disregarded, and the focus is on the team-level variance.

Although from this it may seem that data aggregation is a methodological or statistical issue, it is fundamentally a theoretical one. Within a multilevel framework, data aggregation is about bottom-up processes or the micro-to-macro problem, because it addresses the relation between individual (micro) level observations and team (meso or macro) level constructs and outcomes (Chan, 1998). When aggregating a variable from the individual to the team level, researchers (implicitly or explicitly) make theoretical assumptions about the nature of the aggregated construct, how team level constructs emerge during team interaction, or how individual characteristics or behaviors relate to team processes and outcomes. Which assumptions are relevant depends on the type of variable that is aggregated, as we discuss in the “type of constructs” section below.

Whether data aggregation is appropriate depends on the level at which variables are conceptualized theoretically, and Table 1 presents examples of variables that are often conceptualized at the individual and team level. There are two reasons to aggregate variables. The first is that the specific variable itself is conceptualized at the team level, but can only (or better) be measured at the individual level (e.g., cohesion as a consequence of inter-individual attraction within teams, age diversity of a team). A variable should be conceptualized at the team level when theoretical interest is in variance between teams, and when individual-level data do not capture the essence of the team-level construct. For example, a researcher may not be interested in individual-level perceptions of cohesion (but rather in team-level cohesion as a shared perception), and individual-level age is not a proper measure of age diversity.

The second reason to aggregate is that the phenomenon of interest is at the team level, but some variables were measured at the individual level.

Table 1. Conceptualization of different types of variables

Type of variable	Level at which variable is conceptualized	
	Individual	Team
1. Stable characteristics	Individual difference variables (personality, general mental ability, home situation, demographics)	Global team properties (team size, team location, team function, team type) Team composition <i>Formative team constructs</i> (personality composition, demographic composition, faultlines)
2. Contextual variables	Individual level contextual variables (experienced time pressure, LMX, individual autonomy)	Team level contextual variables <i>Reflective team constructs</i> (leadership style, team autonomy, interdependence)
3. States	Psychological states (mood, motivation, mental model, self-efficacy, job satisfaction, attitudes)	Emergent states ^a : <i>Reflective team constructs</i> (team cohesion, team climate, group affective tone, team potency) <i>Formative team constructs</i> (informal hierarchy, team mental models, network density)
4. Behaviors	Individual behaviors (asking for help, providing information, effort)	Team processes ^a : <i>Reflective team constructs</i> (collective problem solving, collective information sharing) <i>Formative team constructs</i> (conflict handling patterns, information sharing asymmetry)
5. Output	Individual output (performance, creativity, decision quality)	Team output (team performance, team creativity, team decision quality)

^asee Marks et al. (2001)

This situation can occur when the criterion (dependent) variable is conceptualized at the team level. An example is predicting the time a team takes to solve a challenging puzzle (team-level criterion) using individual-level intelligence scores (individual-level predictor). This second situation has broader implications than the first, because it may imply that *all* variables that were measured at the individual level have to be aggregated to the team level, even when they are not conceptualized at the team level (e.g., intelligence is conceptualized at the individual level). The first question in our flow diagram (Figure 1) therefore is: At which level is the criterion variable conceptualized? In cases where criterion variables are conceptualized at the individual level, data should be analyzed at the individual level. When the criterion variable is conceptualized at the team level, data analysis should be done at the team level, and aggregating data to the team level is often needed and appropriate.

Individual level criterion variables

If a researcher is interested in an individual-level construct that is measured in a team context, data aggregation of that construct is not needed, but the team context should still be accounted for. In this case, there is (theoretically meaningful) variation among members of the same team. However, observations of individual-level criterion variables in team contexts are usually non-independent (e.g., Kenny et al., 2002), and not taking this into account during data analysis violates the assumption of independence of observations, which can seriously bias results (e.g., Kenny et al., 2002). The solution is to use multilevel modeling (MLM), which is an extension of the general linear model to analyze data at multiple levels (see e.g., Bliese, 2002; Snijders & Bosker, 1999).

Guideline 1: When a criterion variable is conceptualized at the individual level, but there are multiple observations of individuals in teams, analyze data at the individual level using MLM. Data aggregation in this case is only needed for other variables (e.g., predictors) that are conceptualized at the team level, but measured at the individual level.

Team level criterion variables

Criterion variables that are often conceptualized at the team level include emergent states (e.g., team cohesion, team climate, informal hierarchy, team mental models), team processes (e.g., collective problem solving, cooperative behavior, collective conflict management behaviors), and team output (e.g., team performance; Table 1).¹ Criterion variables that are conceptualized at the team level are frequently also measured at the team level. For example, team performance data might be obtained from sources outside the team (e.g., company records, ratings of the team by outside experts) or from expert sources inside the team (e.g., supervisor ratings). If all members of a team receive the same value on the criterion variable, MLM cannot be applied as there only is team-level variance (unless there is another layer of clustering, such as teams nested in organizations; e.g., Snijders & Bosker, 1999), and data must be analyzed at the team level:

Guideline 2: When criterion variables are conceptualized and measured at the team level, analyze the data at the team level.

Sometimes, however, criterion variables are conceptualized at the team level, but measurement takes place at the individual level (e.g., team cohesion; see also Table 1). Often, the reason is that these variables cannot be directly or unambiguously observed (e.g., by outside raters or team leaders), making it necessary to rely on individual-level observations or ratings. When this is the case, the criterion variable needs to be aggregated to the team level, and analysis should be conducted at the team

level. The reason is that an individual-level variable will not capture what the researcher aims to capture, and analyzing data at the individual level may lead to invalid conclusions.

Two examples are illustrative. First, suppose a researcher wants to examine team cohesion as a team-level construct, but has measured it by asking team members to report on their own experienced team cohesion. Because individual-level variance is not of interest (it reflects differences in individual perceptions rather than differences in team-level cohesion), data analysis at the individual level is not appropriate and may lead to invalid conclusions (e.g., Type I errors).² Second, suppose a researcher is interested in the determinants of team hierarchy, but has measured individual-level influence in the team. Clearly, individual-level influence is an invalid operationalization of team-level hierarchy, because hierarchy implies a *pattern* of influence of multiple members within the team. Computing a team-level measure of hierarchy out of individual-level observations of influence would constitute data aggregation, which would be the appropriate way to proceed.

Guideline 3: When criterion variables are conceptualized at the team level but measured at the individual level, aggregate them and analyze data at the team level.

This also applies to *predictor* variables that were measured at the individual level. When data is best analyzed at the team level (because the criterion variable is theoretically only meaningful at the team level), the next question in our flow diagram consequently focuses on at which level criterion and predictor variables have been measured (Figure 1). When predictors were measured at the individual level whereas the criterion was conceptualized at the team level, aggregating predictors to the team level is both necessary and appropriate. This applies to *all* predictor variables, including variables that are conceptualized at the individual level (e.g., individual-difference measures, psychological states) and at the team level

(e.g., emergent states, team processes; see Table 1). In sum:

Guideline 4: When theoretical interest is at the team level (because the criterion variable is conceptualized at the team level), aggregate predictors that have been measured at the individual level to the team level.

This situation is very common. In the 110 team studies of our sample, 92 had a criterion variable at the team level (75 only at the team level and 17 at both the individual and the team level). It should also be noted that Guidelines 2 to 4 are non-trivial. In our experience as reviewers and authors, it is not uncommon that 1) authors analyze a team-level variable (e.g., team performance) at the individual level (e.g., as individual-level performance perceptions), but still draw (potentially invalid) conclusions at the team level, or that 2) reviewers or editors insist that MLM should be applied in cases in which the phenomenon of interest (the criterion variable) is at the team level.

Types of team constructs

Certain variables do not require aggregation to the team level, because they can be measured at that level. This is true for what Klein and Kozlowski (2000) call “global team properties” (see Table 1), which are conceptualized at the team level, only exist at that level, and are usually considered to be exogenous and relatively stable over time. These variables can often be established unambiguously, or can be established using an expert source (e.g., a team leader, an HR manager). Examples include team size, team location, team type, or company type. Other variables, such as team performance, may not be stable over time and may be endogenous to the research model (e.g., as the criterion variable), but also only exist at the team level and are often also measured at that level. We will not further discuss these variables, because they do not require aggregation.

For variables that do require aggregation to the team level, we draw on a theoretical distinction in psychometrics between *reflective* and *formative*

latent variable models (Bollen & Lennox, 1991; Borsboom et al., 2003; Croon & Van Veldhoven, 2007; Lüdtke et al., 2008) and distinguish between reflective and formative team constructs.³ Our intention is not to go into great depth with regards to the underlying assumptions and psychometric properties of these models. Rather, we use the distinction to clarify a critical theoretical difference between these two types of team-level constructs and to derive implications for proper conceptualization, operationalization, and analysis (examples of both types of team constructs are in Table 1). It is important to note that the distinction between reflective and formative constructs is at the conceptual level and concerns how variables are *theorized*, and not how they have been measured or what are their empirical or statistical properties (although there are implications for measurement and analysis).

Reflective team constructs

A large class of team-level constructs exhibit strong parallels with the logic of the reflective latent variable model. In the reflective latent variable model, theoretical constructs are conceptualized as unobservable (latent) variables that *causally* impact observable indicators in a top-down fashion. This means that the indicators are theoretically interchangeable (Junker & Ellis, 1997) and the information they provide is, in principle, redundant. Moreover, indicators should change in an identical fashion when the latent variable changes, provided that there is no measurement error (which is of course a scientific utopia). Reflective latent variables are assessed through multiple observable indicators (J. R. Edwards & Bagozzi, 2000), where each indicator is a function of the latent variable (Jöreskog, 1971). Moreover, as all indicators change in an identical fashion when the latent construct changes, they are positively correlated (Holland & Rosenbaum, 1986; Krijnen, 2004). Using multiple indicators to assess the latent variable would increase the reliability of the measure when there is measurement

error, but it does not change the nature of the latent variable.

Applied to the current problem, assume a researcher is interested in using team climate to predict team performance. Because team climate is difficult to observe directly, the researcher decides to ask individual members of the team how they perceive the team climate, and to combine their individual responses to estimate team climate. Under the logic of the reflective latent variable model, the latent variable team climate is theorized as *causing* individual team member's responses to questions about the team climate, which are the observable indicators. If there were no measurement error, individual ratings of team climate should be interchangeable and perfectly correlated. Given that measurement error is inevitable, acquiring ratings from multiple members of the team allows the researcher to obtain a more reliable estimate of the latent construct of team climate. For this, a considerable degree of agreement among team members is required for construct validity, because lack of agreement would reflect undesirable levels of measurement error. We propose the term "reflective team constructs" to refer to team-level variables that can be operationalized according to the logic of the reflective latent variable model (Klein & Kozlowski, 2000, call these variables "shared team properties").

It is important to consider what it means when researchers theorize that a variable is a reflective team construct (see also Waller et al., 2016). It implies that there exists a (latent; not necessarily directly observable) property at the team level that *can* actually be experienced and *is* experienced by different members *in a similar way* (e.g., team climate, time pressure, or leadership; see also Borsboom et al., 2003). Because it is assumed that members do not differ substantially in their experiences (although there may be individual level measurement error), there should be agreement among members. Further, each member should be capable of providing valid information on the construct. A good approximation of the "true" value of the construct can therefore be achieved by combining scores of

individuals by computing a team mean (but see our discussion of the "conceptual mean" later on).

Formative team constructs

Another category of team-level constructs shows parallels with the logic of the formative latent variable model. Rather than the latent variable causing its components, in the formative latent variable model the latent variable is *constituted by* (or constructed from) its components so that changes in the components cause changes in the latent variable (Borsboom et al., 2003). In the formative latent variable model, the latent variable is a function of all indicators (Bollen & Lennox, 1991). The latent variable represents an integrated representation of the individual components and is constructed from the components (Borsboom et al., 2003), but is more than a mere composite of them (Bollen & Lennox, 1991). Moreover, indicators do not have to be correlated and are not interchangeable; removing or replacing an indicator will change the constitution of the latent variable.

Applied to the current problem, assume a researcher is interested in using team personality diversity to predict team performance. Because team personality diversity is difficult to observe directly, the researcher decides to measure individual members' personality, and takes the standard deviation of their individual responses to estimate team personality diversity. Under the logic of the formative latent variable model, the latent variable team personality diversity is conceptualized as being constituted by the individual team members' personalities. Importantly, in this case there is no theoretical reason to expect agreement among team members in terms of their personality scores – in fact, variations in agreement (diversity) constitute the variable of interest. This means that team personality diversity is a meaningful and valid construct regardless of how the individual personalities of the team members are related. We propose the term "formative team constructs" to refer to team-level variables that can be operationalized according to the logic

of the formative latent variable model (Klein & Kozlowski, 2000, call these variables “configural team properties”).

When researchers theorize about formative team constructs, the assumption is *not* necessarily that there is a team-level property that is experienced by members in a similar way. In contrast, the property is something that is constructed *by the researcher* based on the aggregation (or configuration) of individual properties or experiences. For example, although team members may be aware of personality diversity in their team, this awareness is *not* a requirement for construct validity when personality diversity is conceptualized as a formative team construct. As a consequence, members do not have to agree on their experiences, are not interchangeable (because no sharedness in experience is assumed), and the team mean is not necessarily the best estimate of the “true” score on a formative latent variable. Instead, how the researcher aggregates formative constructs depends on the theoretical definition of the construct, and on other theoretical considerations, which we will detail in the section on “aggregating formative team constructs”.

Implications

Both types of constructs are very common in team research: of the 301 variables in our sample, 194 were reflective and 107 were formative team constructs. It is, however, likely that they are often used to capture different types of variables. We therefore coded all aggregated variables into the categories of Table 1 (team composition, emergent states, team processes, contextual variables, and team output), and examined whether the type of variable differed between reflective and formative team constructs. This was clearly the case, $\chi^2(4) = 101.27, p < .001$. As is shown in Table 2, reflective constructs were often team processes, emergent states, or contextual variables, whereas formative constructs were often related to team composition and (to a lesser extent) team processes. This makes sense theoretically, given that team members may often share their experience of an emergent state (e.g., cohesion),

Table 2. Types of variables that were aggregated in team studies published in prominent journals between 2016 and 2020 (frequencies and column percentages)

Type of variable	Reflective team constructs	Formative team constructs
Team composition	10 (5%)	57 (53%)
Team contextual variables	50 (26%)	13 (12%)
Emergent states	62 (32%)	8 (8%)
Team processes	67 (35%)	23 (22%)
Team output	5 (3%)	6 (6%)

a team process (e.g., frequency of communication), or a contextual variable (e.g., experienced time pressure). In contrast, group composition often reflects a configuration of individual-level properties (e.g., gender composition), and team processes can be conceptualized as a particular configuration of individual behaviors, such as the degree to which leadership behavior is performed by different team members (i.e., shared leadership; D’Innocenzo et al., 2016).

Because the conceptualization of a team-level variable as a reflective or formative team construct has important consequences, the third question in our flow diagram (Figure 1) is whether the team-level variable is conceptualized as a reflective team construct. One important consequence is whether agreement among team members is required or not. From the previous discussion it becomes clear that such within-team agreement is necessary when and *only* when a variable is a reflective team construct (e.g., Chen et al., 2004; Klein & Kozlowski, 2000). The measurement of a reflective team construct will only be reliable and valid when members of the team show sufficient agreement, and a lack of agreement raises questions about the validity of the measurement of the team-level construct. Agreement is *not* necessary (and often not expected) in the case of formative team constructs.

Guideline 5: When a team-level variable is a reflective team construct, sufficient agreement among team members is required for validity of the construct. This agreement is not required for formative team constructs.

Where (dis)agreement Comes From: team composition, exposure, and emergence

Data aggregation is at its core a theoretical issue concerning the relation between lower-level observations and higher-level constructs and phenomena, and this relation is different for reflective and formative team constructs. When researchers decide on the conceptualization of a team-level variable, they make theoretical assumptions about the processes through which these constructs come into existence and about the nature of these constructs. We recommend that researchers explicitly discuss this, and consider where possible (dis)agreement comes from. We distinguish three situations: (dis)agreement as preexisting, as a consequence of exposure, or as a consequence of emergence.

(Dis)agreement as preexisting. Sometimes, similarities or differences among team members are associated with individual-level variables that are not influenced by the team context. These preexisting variables usually refer to (stable and exogenous) individual-difference variables, such as educational or professional background, personality, abilities or skills, and demographics (see Table 1). Team members may be similar or different on these variables as a consequence of team composition, which can happen either coincidentally (i.e., members happen to be in a team together) or by design (e.g., in a functional versus cross-functional team). Because individual differences are never caused in a top-down way by a latent, team-level variable, team composition (as a direct aggregation of individual-difference variables) should always be treated as a formative team construct. This is also true when managers deliberately assign individuals to teams based

on their individual-level properties, because even then these properties are not a function of a latent variable at the team level (e.g., people do not change their personality when they are assigned to a specific team).

(Dis)agreement and exposure. A second way in which (dis)similarity may arise is through exposure to (tangible or observable) situations or events. For example, members of the same team are often exposed to similar working conditions, leadership, or prior task success, or to similar behaviors from fellow group members. These (team-level) similarities in exposure may, in turn, create shared perceptions of team members about, for example, team-level working pressure, the leadership style of the team leader, or team processes. As a consequence, these types of contextual variables and team process variables are often conceptualized as reflective team constructs (see Table 1 and Table 2). In these cases, it is assumed that some actual situation or event at the team level (e.g., team leader or member actual and observable behavior) *causes* similar experiences or perceptions in team members in a top-down fashion.

Exposure does, however, not always lead to shared perceptions, because members are not always exposed to the same situations or events or interpret or experience them in a different way (e.g., due to individual differences or differences in their role or position in the team). This can create disagreement among members, which can either be conceived as measurement error (not reflecting the actual situation or event) or as meaningful and theoretically interesting variance. In the former case, the construct is still conceptualized as a reflective team construct, but in the latter it is not. In the latter case, Chan (1998) talks about dispersion models, for which the meaning of the construct lies not in the agreement but in the *disagreement* (or dispersion) among team members.

A good example is the construct of LMX differentiation (e.g., Martin et al., 2018). In leader-

member exchange (LMX) theory, leaders are not presumed to treat all team members in a similar way, but to differentiate between ingroup members with whom they have a good and personal relation and outgroup members with whom the relation is less close and more formal (e.g., Graen & Uhl-Bien, 1995). As a consequence, different members will *not* be exposed to the same leader behavior, and will likely *not* develop a shared perception of leadership. The concept of LMX differentiation captures these differences in experience, and is usually operationalized as the within-team variance in individual-level LMX ratings. It is a formative team construct, because the construct is defined in a bottom-up manner as the configuration of individual-level experiences of LMX.

(Dis)agreement as emergence. A third way in which (dis)similarity may arise is through emergence, which refers to phenomena that “exist at a higher or global level, and emerge from the dynamic interactions that take place among components of an underlying level” (Waller et al., 2016, p. 569). In the case of team research, this usually refers to emergent states (see Table 1), which are “properties of the team that are typically dynamic in nature and vary as a function of team context, inputs, processes, and outcomes” (Marks et al., 2001, p. 357). Emergent states are global (they are a property of the team, not of individuals), coherent (they exist for some time, although they can change), ostensive (team members can recognize and experience them), and radically new (not reducible to the lower level; Waller et al., 2016). Even though it is assumed that team members experience them, emergent states can usually not be directly observed (i.e., they are latent). Among the most commonly examined emergent states are team cognition (e.g., shared mental models, transactive memory systems), team climate, team cohesion, and team identification (Rapp et al., 2021).

Although emergent states develop bottom-up out of interactions among team members, once

these latent variables are formed they are often assumed to influence those members in a top-down fashion, making them reflective team constructs. For example, team trust may develop when members note from their interactions that vulnerable people are not exploited in their team, but once it has developed it is supposed to be a real (although not directly observable) attribute of the team that exerts its influence on team members (e.g., they are more willing to speak up). For this assumption to hold, it is thus essential that members experience this state and do so in a similar way (i.e., they agree in their assessment).

However, emergent states can also be formative team constructs (see Table 1). First, as with exposure, not all members may have the same experience in the team, and dispersion may occur. In the case of team climate, dispersion is sometimes referred to as climate strength: Climates are stronger when everyone experiences them in the same way (Chan, 1998), and stronger climates may have stronger effects than weaker climates (Schneider et al., 2002). Climate strength is a formative team construct that focuses on the degree of agreement among members in their (individual) experience of team climate. Another example is shared mental models, which refers to the degree to which team members have a similar (and accurate) understanding of their task or their team (e.g., Mathieu et al., 2000). The degree of sharedness is often of interest, under the assumption that sharedness in a mental model facilitates team coordination (e.g., Lim & Klein, 2006). This conceptualization of mental models makes it a formative team construct, because it is formed by the configuration of (e.g., amount of overlap in) individual mental models (even though, through interaction, members may develop similarities in their mental models; see Edwards et al., 2006).

Second, also differentiation (rather than similarity) can emerge out of interactions among team members. Through interaction, team members can obtain a specific role or position in a team that is different from the role or position of others (e.g., through role-negotiations;

Moreland & Levine, 1982). This can refer both to horizontal differences (e.g., role differentiation) and to vertical differences (e.g., informal hierarchy; Bunderson & Van der Vegt, 2018). For example, DeRue and Ashford (2010) proposed that hierarchies may dynamically develop out of interactions in which members claim or grant leadership or followership roles. The resulting hierarchy (or role distinction between leaders and followers) constitutes a formative construct that may, for example, represent the distribution of influence and power within the team.⁴

In sum, the conceptualization of team-level variables as reflective or formative is based on theoretical considerations of how similarity or agreement versus dissimilarity and dispersion develop. To achieve conceptual clarity about team-level variables, we advise:

Guideline 6: For team-level variables, clearly indicate what type of construct it is (reflective or formative), and why it is conceptualized in this way based on theoretical assumptions about pre-existing differences/similarities, exposure, or emergence.

Aggregation of reflective team constructs

Direct consensus versus referent shift consensus

From theory regarding reflective team constructs, it is clear that agreement (or consensus) among team members on their experience of this construct is important. In this regard, Chan (1998) proposed two different consensus models that are built on different theoretical assumptions: the direct consensus and the referent-shift consensus model (Chan, 1998). In the direct consensus model, the meaning of a higher (team) level construct lies in the consensus (or agreement) among lower-level units (team members), but the construct also has meaning at the individual level. The referent-shift consensus model is similar, but there is a shift in the referent prior to assessment: The

referent is not the individual team member, but the team. In this case variation at the individual level is typically not meaningful but is random error (e.g., due to biases in perception).⁵

An illustrative example is team mood or “group affective tone” (George, 1990). To assess team mood, a researcher may ask all team members to directly indicate their *own* mood state (direct consensus model). The researcher may then aggregate this measure to obtain a measure of “team mood,” but construct validity of this team-level variable would demand that members of the same team overlap in their experienced mood state. Alternatively, researchers may ask team members to report directly on *team* mood, for example through items such as “members of this team are usually in a good mood” (referent-shift consensus model). Also in this case, members should agree in their ratings, because the researcher would assume that individual scores reflect a team-level construct. In our sample of studies, of the 194 reflective team constructs, 41 were measured using the direct consensus model and 138 were measured using the referent shift consensus model (for 15 this was unclear).

Which model is used is important for two reasons. First, a direct consensus procedure may yield a team-level construct that is different from that same construct measured through referent-shift consensus. For example, Van Mierlo et al. (2009) have shown that this is the case for the construct of team autonomy: Direct consensus items (e.g., “Can you influence your work pace?”) measured a different construct than referent-shift items (e.g., “Can your team influence its work pace?”). Second, type of measurement may influence the level of agreement among members. Evidence suggests that levels of agreement tend to be higher when a referent-shift rather than a direct consensus operationalization is used (Arthur et al., 2007; Klein et al., 2001; Van Mierlo et al., 2009). Indeed, one should generally expect considerable agreement among team members when they reflect on the same

team-level phenomenon, more so than when they reflect on their individual experiences.⁶

The primary argument for choosing a direct consensus versus referent-shift model is theoretical. If a researcher assumes that a construct at the individual level is isomorphic to a similar construct at the team level, and if the researcher is interested in this construct at *both* the individual and the team level, or is interested in the “average group member,” then it makes sense to use a direct consensus model. For example, measuring team members’ own mood state would allow one to test whether emotional contagion exists (i.e., that members catch each other’s mood state, implying emergence; see Barsade, 2002) and whether this leads to a shared mood at the team level (see also George, 1990). If, however, a construct is conceptualized exclusively at the team level, and individual-level variation is not of interest, then a referent-shift model is more appropriate: It more clearly captures the team construct, and agreement among team members is likely to be higher.

Guideline 7: When only team-level variance of a reflective team construct is of interest, using a referent-shift model is in most cases preferable. When also individual-level variance is of interest, using a direct consensus model is appropriate.

Measures of agreement

Based on the logic of the reflective latent variable model, sufficient agreement is required for construct validity of a reflective team construct. For continuous measures, this can be established using agreement indices such as the r_{wg} (within-group correlation), ICC(1), and ICC(2) (Intra Class Correlation; see Woehr et al., 2015). The r_{wg} (James et al., 1984, 1993) is a direct measure of agreement (similarity in responses) of members within a team (as compared to random responses), and is computed for each team separately. ICC values, in contrast, are computed across the whole

sample of teams, and do not only assess whether there is agreement *within* a team, but also whether there are differences *between* teams. The ICC(1) provides an estimate of the extent to which ratings of individual members can be attributed to team membership, and represents the proportion of variance at the individual level accounted for by group membership (LeBreton & Senter, 2008). It is possible to compute its significance against the null-hypothesis that it is equal to zero. The ICC(2) indicates how reliably the mean ratings of team members distinguish between teams (Bliese, 2000; Hofmann, 2008). The ICC(1) and ICC(2) are related to each other as a function of team size (Bliese, 1998): When teams are larger, the same ICC(1) will result in a higher ICC(2).

Rather than providing a full discussion of these agreement indices (see LeBreton & Senter, 2008), Table 3 provides some further details on these measures and how to interpret their values. We note that low average values on the r_{wg} index will signal that agreement within teams may be insufficient to justify its conceptualization as a reflective team construct (also see next section). It implies either that measurement error is unacceptably high, or that the theoretical assumption that team members experience the construct in a similar way is not met. Low ICC values, however, can occur when either agreement within teams is low, when differences between teams are low, or both. Low ICC values may therefore also signal that teams cannot be sufficiently distinguished from one another (even despite high agreement within teams). This is problematic from a theoretical point of view, because researchers are generally interested in variance at the team level for reflective team constructs. Moreover, low ICC values will reduce the statistical power of team level models because of restriction of range issues (Bliese, 1998). Given the importance of agreement for reflective team constructs, we recommend:

Table 3. Common indices of intra-team agreement for reflective team constructs and guidelines for their interpretation

Index	Definition	Interpretation (based on LeBreton & Senter, 2008)
r_{wg} (for a single item), and $r_{wg(j)}$ (for scales with more than one item)	The r_{wg} is a measure of the observed distribution of scores within a single team as compared to the expected distribution when there is a complete lack of agreement within a team.	<.30: Lack of agreement .31–.50: weak agreement .51–.70: moderate agreement .71–.90: strong agreement >.91: very strong agreement In published team research, r_{wg} values tend to be high ($M = .88$ in our sample).
ICC(1)	The extent to which ratings of individual members can be attributed to team membership; the proportion of individual-level variance accounted for by group membership.	Interpret like an effect size: .01: small effect .10: moderate effect .25: large effect Typical values in team research are around .20-.25 ($M = .21$ in Woehr et al., 2015; $M = .25$ in our sample).
ICC(2)	How reliably the mean ratings of team members distinguish between teams; the extent to which a mean rating assigned by a group of judges (i.e., team members) is consistent or reliable.	Ideally higher than .60/.70. Typical values in team research are lower ($M = .58$ in our sample). This is because team sizes are often modest, which negatively affects the ICC(2). Lower ICC(2) values may not be problematic for construct validity, but may harm statistical power in the team-level model.

Guideline 8: For reflective team constructs, report values of the ICC(1) and its level of significance and values of the ICC(2) besides r_{wg} coefficients. Discuss how these values compare to previous research and what the consequences are of relatively low values.

Given this guideline, it is important to maximize agreement and reliability of reflective team constructs, which requires careful theorizing and attention to measurement during study design. For example, it is advised to use validated team-level measures of reflective team constructs, and using referent-shift operationalizations (rather than direct consensus) may also increase agreement. Further, extending the sample to include a more diverse collection of teams will increase the amount of variance in individual ratings that can be accounted for by team membership (i.e., higher ICC(1) values).

Getting a high response rate for team members will generally increase ICC(2) values.

Dealing with low agreement when high agreement was expected

In some cases one may end up with low levels of agreement or reliability despite efforts to increase agreement in the study design phase. It is important to consider what to do in such cases, an issue that has so far received only scant attention (but see Chan, 1998; Cole et al., 2011; LeBreton & Senter, 2008). Low agreement may signal unacceptable levels of measurement error, but may also indicate that the theoretical assumptions underlying the conceptualization of a reflective team construct need to be revised.

From a purely methodological point of view, one response would be to conclude that the

measurement of the reflective team construct is invalid and to drop this measure from all analyses. This response is predicated on the assumption that variance at the individual level reflects error variance that is not of interest. This option is not only unattractive from the point of view of wasting resources but, as we argue below, it is also often misguided. Another response is to identify teams that exhibit particularly low levels of agreement (as indexed by the r_{wg}) and to drop them from the sample. For example, Chuang et al. (2016) removed teams that had an r_{wg} value smaller than .50. This may be a viable solution when few teams have (extremely) low agreement, whereas most teams have substantial agreement and when the researcher is interested in the “typical” team (rather than special cases). However, this procedure has notable downsides. First, by dropping data points, potentially meaningful variation is disregarded. Second, dropping teams with low agreement does not necessarily increase statistical power for subsequent hypothesis testing, and may even reduce it because of reduced sample size (Biemann & Heidemeier, 2012; Cole et al., 2011). Third, and most pertinent to our present argument, varying levels of agreement across teams may hold insightful information that is lost when low-agreement teams are dropped from the sample.

From a theoretical point of view, we propose that if proper attention has been devoted to the conceptualization and operationalization of study variables, low agreement does not constitute a “fatal flaw.” Rather, it should be taken as a prompt to further explore the potential sources and nature of the suboptimal agreement as well as its implications for the theoretical understanding of the phenomenon of interest. When some teams exhibit low agreement, this may signal that another conceptualization of the variable is required (e.g., Chan, 1998), because certain theoretical assumptions about this variable as a reflective team construct are not met. We therefore recommend that researchers examine the observed (and unexpected) lack of agreement more closely, and proceed according to the following three steps.

The first step is to look more closely within teams at how responses are distributed. Low agreement may result from members seemingly giving random responses (lack of agreement), but may also stem from a bi-modal distribution with some members giving high scores and others low scores (complete *disagreement*). Such a bi-modal distribution might be evidence for the existence of subgroups within the larger team, and this may require a re-conceptualization of the team variable (i.e., it is not a property of teams, but of subgroups). Another possibility is that one team member has a score that differs greatly (e.g., is much lower) from those of other team members, which may have large effects on r_{wg} values. In the case of psychological safety, for example, this may indicate that this person is socially excluded from the team, which provides information about within-team dynamics. In both cases, it is apparent that the assumption that all members experience a reflective team construct in the same way is not met.

The second step focuses on why agreement in some teams is higher than in other teams. In this step, team-level predictor variables (such as global team properties or team composition) are used to predict (at the team level) the level of agreement of teams (e.g., using the standard deviation as a criterion variable). It is conceivable that agreement for a reflective team construct depends, for example, on team longevity, team size, or team diversity (with higher agreement in teams that have been together longer, are smaller, or are less diverse; e.g., Carter et al., 2018). Evidence of this kind gives information about how agreement develops over time and perhaps about why agreement is low (e.g., most teams in the sample have low team tenure or are relatively diverse).

The third step is predicting individual-level scores on the reflective team construct in an effort to understand why team members experience the construct in different ways. Per team, this involves computing the deviation from the team mean for each team member, and analyzing why some team members deviate more

from the team mean than others. Predictors are at the individual rather than the team level, and can include stable individual differences, data on an individual's position in a team (e.g., status as a newcomer), or cross-level interactions between individual and team-level variables. For example, the only woman in an otherwise all-male team may give a different response than the men. This would show up as an interaction between gender (at the individual level) and group composition (e.g., proportion of men, at the team level) on individual responses. It is recommended to use MLM for such analyses.

Depending on what one concludes from these explorations, it may be insightful to make the level of agreement an integral part of the research model (i.e., as a dispersion model; Chan, 1998). For instance, one may include level of agreement as a predictor and allow it to interact with other predictors to see whether their relationships with some criterion are specific to high-agreement teams (see Cole et al., 2011, for guidelines). This would imply that the level of agreement constitutes a formative team construct, and that the researcher would have to give a theoretical conceptualization of this variable (e.g., that it represents climate strength). Such a re-conceptualization implies that within-team variance of a reflective team construct no longer is conceptualized as error, but as potentially meaningful variance.

Guideline 9: For reflective team constructs, ensure high agreement among team members through design. If agreement is insufficient, systematically explore the causes of disagreement and potentially reconceptualize the reflective team construct.

Aggregating reflective team constructs: the conceptual mean

When aggregating reflective team constructs, such as team climate, the aim is to estimate the latent (unobserved) team-level score on the

dimension of interest, which caused (is reflected in) the observed scores at the individual level. Following our theorizing about reflective formative constructs, we call this latent variable the *conceptual mean* (although, strictly speaking, a latent variable is not a mean). In practice, it is common to aggregate reflective team constructs by taking the average score within each team (i.e., the *arithmetic mean*): Of the 194 reflective team constructs that we coded, 191 were aggregated to the arithmetic mean (98.5%). The conceptual mean, however, is not the same as the arithmetic mean, and the arithmetic mean may not always be a good estimate of the conceptual mean.

There are two problems with aggregating a reflective team construct to the arithmetic mean. First, it implies that one disregards within-team variability and in fact assumes that there is no error variance at the individual level. This assumption only holds with very high ICC(2) values, which are rarely obtained in team research (see Table 3). The more team members are missing in the sample, or the smaller the teams, the more likely it is that the arithmetic mean differs from the conceptual mean (there is too much 'noise' in the measure), which can lead to biased estimates in subsequent analyses (Croon & Van Veldhoven, 2007; Lüdtke et al., 2008). Second, using the team-level arithmetic means might cause variability in the error variances of the criterion variable (i.e., heteroscedasticity), which can both inflate Type I errors and reduce statistical power, leading to erroneous conclusions (Foster-Johnson & Kromrey, 2018; Hayes & Cai, 2007). For example, Croon and Van Veldhoven (2007) showed that this is likely unless team sizes are equal (which is rarely the case, particularly in field research). Therefore, when aggregating a reflective team-level construct the researcher should take steps to counteract these biases. We identified three approaches that remedy (some of) these issues and provide an example R code that implements each of these three methods in Appendix A.

The most straightforward approach consists of taking the arithmetic mean of individual-level measurements, and then applying a heteroscedasticity correction when using this team-level variable in subsequent modeling (Hayes & Cai, 2007; MacKinnon & White, 1985). This correction adjusts standard errors, *t*-values and *p*-values to be more robust against heteroscedasticity. It can substantially increase statistical power in the team-level model (Foster-Johnson & Kromrey, 2018), has few drawbacks (Hayes & Cai, 2007), and can be applied regardless of whether heteroscedasticity is actually present. This approach, however, only corrects inferential bias (e.g., significance levels), but does not provide a better estimate of the conceptual mean, and does not correct bias in (beta) coefficients. The recommended correction is known as HC3, and is available in R (using the sandwich package; Zeileis, 2004), Stata (using the “vce(hc3)” option for the regress command; StataCorp, 2023), and via a freely available macro for SPSS and SAS (published in Hayes & Cai, 2007).

The second approach also involves calculating a new team-level variable and is known as the two-step procedure (Becker et al., 2018; Griffin, 1997). In Step 1, the individual-level measurements are modeled with a random team-level intercept using MLM. The team-level intercepts are estimates of the conceptual mean, and are extracted from the model (e.g., using the ranef function in the lme4 package for R). In Step 2, these team-level intercepts are used in subsequent analyses instead of the arithmetic mean. MLM explicitly separates within-team variance from between-team variance and controls for within-team variance by ‘shrinking’ unreliable team-level intercepts (e.g., as a result of having data from few team members or high within-team variability) towards the grand mean. This produces a more accurate estimate of the conceptual mean compared to calculating the arithmetic mean (Becker et al., 2018). MLM is implemented in all major general-purpose statistical packages.

The third approach directly uses the psychometric theory behind reflective team constructs,

where individual-level measurements reflect within-team variance, a team-level latent mean, and measurement error. This approach estimates individual-level and team-level models simultaneously from the individual-level data (Christ et al., 2017; Lüdtke et al., 2008; Marsh et al., 2009; Preacher et al., 2011) using Multilevel Structural Equation Modeling (MSEM). The conceptual mean is modeled as a latent team-level variable, and the relation between this latent variable and other variables can be investigated in the same model. MSEM is implemented with varying capabilities in most SEM software packages, including Mplus (Muthén & Muthén, 2011), LISREL (Jöreskog & Sörbom, 1993), and the lavaan package for R (Rosseel, 2012).

From this, it is clear that MSEM-type approaches fit best with theory regarding reflective team constructs and are expected to produce the most accurate estimates of the conceptual mean (Lüdtke et al., 2008; Preacher et al., 2011). However, two further considerations should be factored into the decision of which method to use: (a) sample size, and (b) the researcher’s theoretical interest. Regarding sample size, MSEM is the most complicated model, because it estimates parallel models at both the team level and the individual level. Hence the sample size at each level needs to be sufficient, and we recommend at least 50 teams for a simple MSEM model, and 100 teams for more complex models (see Christ et al., 2017, for a further discussion). Regarding theoretical interest, the two-step method has two unique properties compared to the other two. First, it allows the researcher to include individual-level and team-level variables as predictors in Step 1. This allows researchers to model individual-level processes as mediators of either individual-level (the so-called 1-1-2 model; Preacher et al., 2011) or team-level predictors (the 2-1-2 model). A second potential advantage is that MLM can accommodate non-normal dependent variables (e.g., the number of products produced by a team, which follows a Poisson distribution).

Guideline 10: Aggregate reflective team constructs by estimating the conceptual mean. When sample sizes are sufficiently large (> 100 teams), use MSEM; when the research interest is in cross-level processes or when analyzing a non-normal dependent variable, use the two-step approach. In other cases, we recommend using the arithmetic mean and correcting for heteroscedasticity in the team-level model.

These more recent developments in data aggregation have not (yet) been widely applied in the literature. As noted, in most papers the arithmetic mean was used to aggregate reflective team constructs (98.5% of constructs), and correction for heteroscedasticity was never applied. The two-step procedure was never applied either, but MSEM was applied in three cases (Dierdorff et al., 2019; Koopmann et al., 2016; Larson et al., 2020). Although alternative procedures to using the arithmetic mean may sometimes yield comparable results, it also seems that team research is at risk of drawing invalid conclusions.

Aggregation of formative team constructs

Well-Defined and ill-defined formative team constructs

Formative team constructs are measured at the individual level, but inform the development or meaning of a construct at the team level (see Table 1). These variables originate at the individual level, but they are *not* assumed to converge among team members (although they may do so, e.g., as a consequence of attraction-selection-attrition processes). Although agreement is not a critical issue with formative team constructs, operationalization sometimes is: Because no agreement is required or expected, the (conceptual) mean is not necessarily the best way to operationalize a formative team construct (as was the case with reflective team constructs). Rather, the operationalization of formative team constructs is chosen by the researcher and

should be informed by theory about how a team-level construct relates to individual-level observations. The way in which individual characteristics or contributions are captured (i.e., operationalized) at the team level may therefore vary depending on theoretically proposed patterns, distribution, and/or variability among members' contributions to the unit-level phenomenon (Klein & Kozlowski, 2000). Given that there are many different ways in which such variables could be aggregated, it is important to understand which aggregation approach is the most suitable for a specific variable.

To help researchers in deciding on the best aggregation approach, we find it useful to consider the degree to which formative constructs are well-defined at the team level. Well-defined formative constructs have a conceptual team-level meaning, and the construct definition gives clear insight into the appropriate aggregation method, whereas for ill-defined formative constructs this is not the case. Considerations about how to aggregate a formative team construct differ between these two types of construct.

Aggregating well-defined formative team constructs

For well-defined formative constructs the construct definition provides insight into how the team-level construct is made up of the individual-level inputs. For instance, when one is interested in the effect of ethnic diversity on team performance, one should aggregate the ethnicity of the individual team members to a diversity index by means of Blau or Teachman's entropy index (Harrison & Klein, 2007). Team network density, defined as the overall level of interaction reported by team members (Sparrowe et al., 2001), is best operationalized as the sum of the actual individually reported ties divided by the total possible sum of ties. Similarly, the definition of team conflict asymmetry as the degree of "variation, or dispersion, in members' perceptions of the level of conflict in their group" (Jehn et al., 2010, p. 597) directly suggests the use of the standard deviation of team members' perceptions on conflict scores.

The degree to which a construct is well-defined is not necessarily binary, because it also depends on conventions in (or maturity of) the literature (i.e., whether a widely accepted measure is available). For example, the construct of shared leadership is defined as “an emergent and dynamic team phenomenon whereby leadership roles and influence are distributed among team members” (D’Innocenzo et al., 2016, p. 1968). From this definition, it may not be immediately clear how this distribution of leadership roles should be operationalized at the team level. However, in recent work shared leadership has often been approached from a social network perspective, which suggests that density in the leadership network is an appropriate operationalization (e.g., Kukenberger & D’Innocenzo, 2020; Lorinkova & Bartol, 2020). When relying on a social network conceptualization, this would make shared leadership a relatively well-defined formative team construct.

Many formative constructs are associated with diversity (or its flip side, homogeneity or sharedness). Even though diversity seems well-defined, Harrison and Klein (2007) argued that diversity is actually not one thing, but three things – variety, separation, and disparity – that are associated with different operationalizations and map onto different theoretical arguments. What is maximum diversity depends on the specific diversity type. Thus, when one examines variety, a team is most diverse when it shows a uniform distribution, with all team members differing on a categorical variable (e.g., all members hold a different educational degree). Separation is the most extreme when there is a bimodal distribution, with half of the team members scoring low of and the other scoring high on a continuum (e.g., half is completely in favor of and half is completely against a proposal). Finally, when one examines diversity as disparity, diversity is maximal when the team is positively skewed on a continuum (e.g., one team member has absolute power, whereas the other members have no power). How these different types of diversity are linked to team outcomes is informed by different theories: variety effects are linked to information/decision-making processes, separation effects are

informed by similarity/attraction approaches, and disparity effects are explicated in justice theories (Harrison & Klein, 2007). When a researcher is interested in team diversity, it is therefore advisable to specify the type of diversity (variety, separation, or disparity), so the theoretical construct is linked explicitly to its operationalization.

In general, making sure that a formative construct is well-defined links its construct definition better to its operationalization. This has obvious advantages for construct clarity and for consistency of operationalizations across different studies. We therefore advise:

Guideline 11: When formative team constructs are well-defined, the specific aggregation approach is informed by the definition of the construct. If possible, make sure that these constructs are well-defined and include in the construct definition how the team-level construct is made up of the individual-level inputs.

In our sample, we coded 39 formative constructs as well-defined, and 67 as ill-defined (we could not classify one: self-awareness in Dierdorff et al., 2019). Well-defined constructs were often diversity related (fourteen cases) and were aggregated by using Blau’s index (i.e., variety; six cases), within-team standard deviation or similar (separation; six cases), or skewness (disparity; two cases). In 22 other cases, some social network measure was used (e.g., network density for shared leadership). In the other three cases, some other form of aggregation was used, such as faultline strength.

Aggregating ill-defined formative team constructs

For formative constructs that are ill-defined, the link between conceptualization and operationalization of a variable is not clear a priori. We propose that the aggregation method should in this case depend on theory regarding how individual team members determine processes and outcomes of the team. The most commonly

studied formative construct that does not have a clearly defined conceptual meaning at the team level is team personality composition (Barrick et al., 1998; Bell, 2007; Halfhill et al., 2005; LePine, 2003; Neuman et al., 1999; Peeters et al., 2006), but similar arguments apply to team composition based on other types of variables (e.g., tenure, abilities, demographics). In these cases, the team-level construct is not merely a description of the individual members' attributes, but may refer to some configuration of individual-level attributes (Mathieu et al., 2008), and a priori it will not necessarily be clear which type of configuration to focus on (e.g., average, diversity, or extremes). Indeed, of the 67 ill-defined formative constructs, 43 (64%) were related to team composition in some way. The others related to certain configurations of individual-level behaviors, or individual-level performance, for example.

One approach to deciding on an aggregation technique is determining the team members' relative input in team performance (i.e., their interdependence; Barrick et al., 1998; Chan, 1998; LePine et al., 1997). The assumption is that personality (or some other factor) shapes individual input (e.g., members higher in openness to experience provide more creative input), and that the specific form of within-team interdependence determines how individual inputs are translated to team outputs (e.g., only one creative member is needed to make the team creative; "creative stars"; see Li et al., 2020). In Steiner's task taxonomy (Steiner, 1972), tasks are divided in terms of interdependence in additive (where individual effort is summed to get to team performance), compensatory (where individual effort is averaged), disjunctive (where the best performing team member determines team performance), conjunctive (where the worst performing team member determines team performance), and discretionary (where the team decides how they will use or combine individual team members' input). When aggregating team personality to the team level to predict conjunctive task performance, taking the lowest score (i.e., the minimum) on the personality trait is most appropriate.

Alternatively, when the task is compensatory, taking the average of the trait is the most suitable approach. Note, however, that this average does *not* assume a reflective team construct that exerts a top-down influence on team members (e.g., "team personality"), but simply is a specific aggregation of an individual trait.

However, using Steiner's taxonomy may not always be possible (e.g., when interdependence is unclear or mixed), or may not always be the best way to proceed. If the crucial mechanism is not how individual contributions are combined into a team (performance) outcome, but rather into some other team process (e.g., conflict or cooperation), aggregation should be based on this intervening process. For example, if outcomes are assumed to be a consequence of cooperation as shaped by team members' agreeableness, one could aggregate by taking the lowest score on agreeableness in the team, under the assumption that "one bad apple will spoil the bunch." Clearly, such decisions require theory about relevant team processes.

In practice, a majority of ill-defined team constructs in our sample was aggregated by taking the arithmetic mean (60%; 40 out of 67) or related methods (median, count, sum: seven cases). Other popular aggregation methods were diversity-related (e.g., standard deviation; eight cases) or taking the score of one particular team member (e.g., highest scoring, score of a core member; four cases). The other ways of aggregation were more idiosyncratic (e.g., whether or not at least two women were on a team; Graham et al., 2020).

The general conclusion with regards to formative team constructs is that when aggregating such constructs to the team level it is crucial to have strong theory that informs the specific operationalization. Sometimes this theory is an integral part of the conceptualization of the formative team construct, which makes the aggregation approach straightforward (i.e., team conflict asymmetry should be operationalized as variance), but sometimes additional theory about team processes is necessary to understand how non-linear individual inputs are linked to

the team's output (e.g., when aggregating individual team members' personality scores; Homan et al., 2008; Van Kleef et al., 2009). It is important that researchers explicitly consider and justify why a certain aggregation method is used.

Guideline 12: When formative team constructs are ill-defined, base the specific aggregation approach on theory that explains how individual team members' contributions are related to processes within and/or outcomes of the team, and explicitly justify the chosen method of aggregation.

Finally, it may be important to consider more complex combinations of properties of individuals and properties of teams when deciding on an aggregation strategy. For example, researchers may consider the distribution of task roles, an individual's place in the team's hierarchy, and being a critical (or core; vs. peripheral) member of the team. Indeed, certain team members might be relatively more central or critical in the team's social network (e.g., Ellis et al., 2005; Humphrey et al., 2009), and core team members might be more influential in the cooperation and communication processes within the team. Their individual-level properties (e.g., expertise, personality) may therefore be relatively influential in determining team outcomes, perhaps in concert with properties of teams (e.g., team climate, size, or hierarchy). Developing theories along these lines would greatly advance team research.

General discussion and conclusion

Summary of key points

Bridging the levels of analysis in team research can be tricky. Most methodological work has focused on the lower, individual level of analysis, and has addressed how to deal with statistical dependency in data (e.g., by using multilevel modeling; Kenny et al., 2002; Snijders &

Bosker, 1999). Far less work has examined theoretical and methodological issues regarding the higher, team level of analysis. Critical in this respect is the issue of data aggregation, or how to construct team-level variables from individual-level data. Although important papers have appeared over the years (e.g., Chan, 1998; Chen et al., 2004; Klein et al., 1994; Klein & Kozlowski, 2000; LeBreton & Senter, 2008), the different insights have not been sufficiently integrated into a coherent framework that also offers a practical guide for researchers. The main aim of the present paper was to provide such a framework (see Figure 1). This framework will hopefully be a useful guide for researchers, reviewers, and editors, but may also serve to correct common misunderstandings and overgeneralizations (e.g., that multi-level data should *always* be analyzed with multilevel modeling (MLM), or that within-team agreement is *always* a requirement for aggregation).

In integrating the scattered insights in the literature, we introduced new distinctions, concepts, and guidelines to the study of teams. First, and most importantly, we borrowed insights from psychometrics (e.g., Croon & Van Veldhoven, 2007; Lüdtke et al., 2008) by making a theoretical distinction between reflective and formative team constructs. Second, and based on this distinction, we provided three ways through which individual and team constructs are related: as preexisting similarities and differences, and as a result of (shared) exposure or as a result of emergence. Third, we discussed the role of agreement in data aggregation, and provided new guidelines about possible actions that researchers can take when agreement is (too) low. Fourth, for reflective team constructs, we discussed recent developments pointing to alternatives to taking the arithmetic team mean as an operationalization of the construct, and we provided guidelines as to which method should be preferred when. Fifth, and finally, we made a distinction between well-defined and ill-defined formative team constructs, and provided guidelines for the aggregation of both.

Throughout, we illustrated different concepts and ideas with a snapshot of current research practices, which resulted in some interesting observations. First, data aggregation is very common in team research. Second, a wide range of team variables are commonly aggregated and this includes a range of both reflective and formative team constructs. Third, by far the most common way to aggregate variables in team research is taking the arithmetic mean, and this is the case for formative constructs (60%) and overwhelmingly so for reflective team constructs (98.5%). This practice continues, despite the fact that using the arithmetic mean to aggregate reflective team constructs may lead to biased results (e.g., when team sizes are unequal). Fourth, and despite the dominance of mean-based aggregation, a number of other popular aggregation methods exist, most notably diversity indices and (more recently) social network measures (e.g., for shared leadership).

Future research

From a theoretical point of view, a main challenge in team research is to develop strong theory about how individual-level constructs relate to team-level constructs and processes. On the one hand, this applies to the conceptualization of aggregated team-level constructs. As we have emphasized in Guideline 6, researchers should specify what type of construct (reflective or formative) is concerned, and explain why this is the case or which assumptions underlie this conceptualization (e.g., how agreement or disagreement among team members in this variable develops). For formative team constructs in particular, it helps to have a clear conceptualization of the construct that also suggests a way to aggregate it (i.e., develop well-defined constructs; guideline 11). A specific theoretical approach to this is helpful, as is the case in, for example, the social network approach to shared leadership (e.g., Kukenberger & D’Innocenzo, 2020; Lorinkova & Bartol, 2020).

On the other hand, for ill-defined formative constructs, it is important to have clearly

articulated theory on how (aggregated) individual-level variables relate to team-level processes and outcomes. This applies mainly to work that looks at team composition effects, in which one or more individual-difference variables are aggregated to the team level. Unfortunately, we have no generally agreed upon framework for this, although there are some developments. For example, Humphrey et al. (2009) have developed a theory about the strategic core of teams, which suggests that characteristics of core role holders in a team may be more important than characteristics of those that do not hold these roles. Further, Emich et al. (2021) have developed an approach based on attribute alignment, which suggests that researchers should perhaps focus less on team composition in terms of single attributes of members and more on the alignment of attributes. For example, members who are high on both conscientiousness and pro-activity may be particularly effective team players, and teams who have (many of) these members may be more effective.

From a methodology point of view, at least two issues clearly deserve further research attention. First, we noted that aggregating a reflective team construct by taking the arithmetic mean is quite common, but also potentially problematic. In recent years, several procedures have been suggested to correct or prevent possible biases in results and conclusions, including correcting standard errors for heteroscedasticity, using the two-step approach, and multilevel structural equation modeling. What is needed, however, is research in which the performance of these different approaches is compared under different circumstances, such as different sample sizes, varying team sizes, or different levels of reliability (e.g., ICC(2)). Because a systematic investigation is yet to be conducted, our guidelines in this respect are preliminary.

Another issue is how to deal with missing data in team research. It is common that researchers do not have a 100% response rate per team, and that estimates of aggregated team-level constructs are consequently based on a limited sample of team members. From our

sample of papers, it appears that different authors treat this issue differently. Whereas some explicitly noted that they did not exclude teams due to missing data (e.g., Mao et al., 2021), others did exclude teams based on missing data using various cut-off criteria. Two popular exclusion criteria were fewer than three responses per team (six cases) or a minimum response percentage per team (also six cases, although different values were used: 60%, 70%, 75%, and 80%). This raises the question what best practice in this regard would be.

We suggest that the answer to this question depends on the nature of the data and the research model. On the one hand, as noted by Biemann and Heidemeier (2012) and Cole et al. (2011), excluding teams due to low response may not be a good idea, because this reduces statistical power (despite potentially higher reliability in the measurement of team constructs). On the other hand, missing data may lead to invalid estimates of an aggregated variable. This issue is potentially less severe when aggregating a reflective construct with a (very) high ICC(2) (i.e., implying that team members are more or less interchangeable). However, it becomes more severe with lower levels of ICC(2) in the case of reflective team constructs, when missing data is systematic, or with (certain) formative constructs. For example, when aggregating a formative team construct by taking the lowest score in the team, missing data may have a large impact when the person with the lowest score did not respond. Clearly, the issue deserves further study, and for now we can only repeat our advice that it is important to invest in getting high response rates. If, for some reason, this is not possible, we recommend analyzing the data both with and without exclusions based on response rates, and seeing whether and how this influences the results.

Conclusion

Data aggregation in team research is very common, but can also be very tricky. With

this contribution, in which we build on a fundamental distinction between reflective and formative team constructs, we hope to have provided useful guidelines for researchers, editors, and reviewers. This will hopefully prevent misunderstandings and errors and help to move the field of team research forward. Our emphasis on the importance of theoretical groundwork, next to methodological and statistical considerations, can stimulate new research questions, better fitting aggregation techniques, and more reliable team data. Additionally, we hope these insights will inspire researchers to explore and learn from (rather than discard) team data that show low agreement in order to develop a more comprehensive understanding of the rich yet complex dynamics of teams.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Bernard A. Nijstad  <https://orcid.org/0000-0001-7882-8447>

Astrid C. Homan  <https://orcid.org/0000-0002-6795-7494>

Marc W. Heerink  <https://orcid.org/0000-0002-6931-6519>

Supplemental material

Supplemental material for this article is available online.

Notes

1. Emergent states are defined as “properties of the team that are typically dynamic in nature and vary as a function of team context, inputs, processes, and outcomes” (Marks et al., 2001, p. 357). Team processes are defined as “members’ interdependent acts that convert inputs to outcomes through cognitive, verbal, and behavioral activities directed towards organizing taskwork to achieve collective goals” (Marks et al., 2001, p. 357).
2. Doing so would constitute an atomistic fallacy: Conclusions at the individual level of analysis do not necessarily generalize to the team level of

analysis (the level one is interested in). Similarly, conclusions at the team level do not necessarily generalize to the individual level, and inappropriately drawing conclusions about individuals from team-level results is called the ecological fallacy (see e.g., Hannan, 1971; Robnson, 1950; Thorndike, 1939). In some cases, the sign of the correlation between two variables (positive vs. negative) may even be different at different levels of analysis.

3. Klein and Kozlowski (2000) use the distinction between “shared team properties” (similar to reflective team constructs) and “configural team properties” (similar to formative team constructs). We prefer the labels reflective and formative for two reasons. First, the label “shared team properties” may suggest that one has empirically established sharedness in that property. With the label reflective team construct we emphasize how the variable is conceptualized, whether sharedness is empirically observed or not. Second, these labels clearly indicate a relation with the logic of latent variable models, and are more closely aligned with recent developments in the aggregation of reflective team constructs, which we discuss in the section on the conceptual mean. The distinction between reflective and formative constructs has also been made by others (Croon & Van Veldhoven, 2007; Lüdtke et al., 2008).
4. Sometimes, researchers measure team constructs that we label as formative in Table 1, such as team diversity or hierarchy, in such a way that they are effectively turned into reflective team constructs. In these cases, team members are directly asked to assess, for example, team diversity (“how diverse is your team?”) or hierarchy (“how hierarchical is your team?”), and it is assumed that team members will agree in this assessment (e.g., because of shared exposure).
5. An exception is when a researcher is explicitly focused on divergent perceptions of team-level phenomena, as we discussed as dispersion models in the previous section.
6. An assumption underlying this is that team members have sufficient insight into the team construct to be able to rate it. At times, this may not be the case, and it may be better to ask team members to reflect on their own personal experiences or behavior, for example when they cannot know or judge the experiences of others in the team. In those cases, a direct consensus measurement may be preferred.

References

- Allen, N. J., Stanley, D. J., Williams, H., & Ross, S. J. (2007). Assessing dissimilarity relations under missing data conditions: Evidence from computer simulations. *Journal of Applied Psychology, 92*(5), 1414–1426. <http://dx.doi.org/10.1037/0021-9010.92.5.1414>
- Arthur, W., 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*(1), 35–58. <https://doi.org/10.1177/1094428106287574>
- Barrick, M. R., Neubert, M. J., Mount, M. K., & Stewart, G. L. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology, 83*(3), 377–391. <https://doi.org/10.1037/0021-9010.83.3.377>
- Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly, 47*(4), 644–675. <https://doi.org/10.2307/3094912>
- Becker, D., Breustedt, W., & Zuber, C. I. (2018). Surpassing Simple Aggregation: Advanced Strategies for Analyzing Contextual-Level Outcomes in Multilevel Models. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (MDA), 12*(2), 233–263. <https://doi.org/10.12758/mda.2017.05>
- Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology, 92*(3), 595–615. <https://doi.org/10.1037/0021-9010.92.3.595>
- Biemann, T., & Heidemeier, H. (2012). Does excluding some groups from research designs improve statistical power? *Small Group Research, 43*(4), 387–409. <https://doi.org/10.1177/1046496412443088>
- Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*(4), 355–373. <https://doi.org/10.1177/109442819814001>
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. *Multilevel theory,*

- research, and methods in organizations (pp. 349–381). Jossey-Bass.
- Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In F. Drazgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 401–445). Jossey Bass.
- 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), 1–23. <http://dx.doi.org/10.1007/s10869-017-9491-z>
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. <https://doi.org/10.1037/0033-295X.110.2.203>
- Bunderson, J. S., & Van der Vegt, G. S. (2018). Diversity and inequality in management teams: A review and integration of research on vertical and horizontal member differences. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*(1), 47–73. <https://doi.org/10.1146/annurev-orgpsych-032117-104500>
- Carter, N. T., Carter, D. R., & DeChurch, L. A. (2018). Implications of observability for the theory and measurement of emergent team phenomena. *Journal of Management*, *44*(4), 1398–1425. <https://doi.org/10.1177/0149206315609402>
- 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*(2), 234–246. <https://doi.org/10.1037/0021-9010.83.2.234>
- Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multi-level construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Multi-level issues in organizational behavior and processes (Research in Multi-Level Issues, Vol. 3)* (pp. 273–303). Emerald. [https://doi.org/10.1016/S1475-9144\(04\)03013-9](https://doi.org/10.1016/S1475-9144(04)03013-9)
- Christ, O., Hewstone, M., Schmid, K., Green, E. G. T., Sarrasin, O., Gollwitzer, M., & Wagner, U. (2017). Advanced multilevel modeling for a science of groups: A short primer on multilevel structural equation modeling. *Group Dynamics*, *21*(3), 121–134. <https://doi.org/10.1037/gdn0000065>
- Chuang, C. H., Jackson, S. E., & Jiang, Y. (2016). Can knowledge-intensive teamwork be managed? Examining the roles of HRM systems, leadership, and tacit knowledge. *Journal of Management*, *42*(2), 524–554. <https://doi.org/10.1177/0149206313478189>
- Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management*, *23*(3), 239–290. <https://doi.org/10.1177/014920639702300303>
- Cole, M. S., Bedeian, A. G., Hirschfeld, R. R., & Vogel, B. (2011). Dispersion-composition models in multilevel research: A data-analytic framework. *Organizational Research Methods*, *14*(4), 718–734. <https://doi.org/10.1177/1094428110389078>
- 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*(1), 45–57. <https://doi.org/10.1037/1082-989X.12.1.45>
- DeRue, D. S., & Ashford, S. J. (2010). Who will lead and who will follow? A social process of leadership identity construction in organizations. *Academy of Management Review*, *35*(4), 627–647. <https://doi.org/10.5465/amr.35.4.zok627>
- Dierdorff, E. C., Fisher, D. M., & Rubin, R. S. (2019). The power of percipience: Consequences of self-awareness in teams on team-level functioning and performance. *Journal of Management*, *45*(7), 2891–2919. <https://doi.org/10.1177/0149206318774622>
- D’Innocenzo, L., Mathieu, J. E., & Kukenberger, M. R. (2016). A meta-analysis of different forms of shared leadership–team performance relations. *Journal of Management*, *42*(7), 1964–1991. <https://doi.org/10.1177/0149206314525205>
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. <https://doi.org/10.1037/1082-989X.5.2.155>

- Edwards, B. D., Day, E. A., Jr, W. A., & Bell, S. T. (2006). *Relationships Among Team Ability Composition, Team Mental Models, and Team Performance*, 91(3), 727–736. <https://doi.org/10.1037/0021-9010.91.3.727>
- Ellis, A. P. J., Bell, B. S., Ployhart, R. E., Hollenbeck, J. R., & Ilgen, D. R. (2005). An evaluation of generic teamwork skills training with action teams: Effects on cognitive and skill-based outcomes. *Personnel Psychology*, 58(3), 641–672. <https://doi.org/10.1111/j.1744-6570.2005.00617.x>
- Emich, K. J., Lu, L., Ferguson, A., Peterson, R. S., & McCourt, M. (2021). Team composition revisited: A team member attribute alignment approach. *Organizational Research Methods*, 25(4), 642–672. <https://doi.org/10.1177/10944281211042388>
- Foster-Johnson, L., & Kromrey, J. D. (2018). Predicting group-level outcome variables: An empirical comparison of analysis strategies. *Behavior Research Methods*, 50(6), 2461–2479. <https://doi.org/10.3758/s13428-018-1025-8>
- George, J. M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology*, 75(2), 107–116. <https://doi.org/10.1037/0021-9010.75.2.107>
- Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *The Leadership Quarterly*, 6(2), 219–247. [https://doi.org/10.1016/1048-9843\(95\)90036-5](https://doi.org/10.1016/1048-9843(95)90036-5)
- Graham, M. E., Walia, B., & Robinson, C. (2020). Women executives and off-the-job misconduct by high-profile employees: A study of National Football League team organizations. *Journal of Organizational Behavior*, 41(9), 815–829. <https://doi.org/10.1002/job.2476>
- Griffin, M. A. (1997). Interaction between individuals and situations: Using HLM procedures to estimate reciprocal relationships. *Journal of Management*, 23(6), 759–773. <https://doi.org/10.1177/014920639702300604>
- Halfhill, T., Sundstrom, E., Lahner, J., Calderone, W., & Nielsen, T. M. (2005). Group personality composition and group effectiveness: An integrative review of empirical research. *Small Group Research*, 36(1), 83–105. <https://doi.org/10.1177/1046496404268538>
- Hannan, M. T. (1971). *Aggregation and disaggregation in sociology*. Heath-Lexington.
- Harrison, D. A., & Klein, K. J. (2007). What's the difference? diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review*, 32(4), 1199–1228. <http://dx.doi.org/10.5465/amr.2007.26586096>
- Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, 39(4), 709–722. <https://doi.org/10.3758/BF03192961>
- Hirschfeld, R. R., Cole, M. S., Bernerth, J. B., & Rizzuto, T. E. (2013). Voluntary survey completion among team members: Implications of non-compliance and missing data for multilevel research. *Journal of Applied Psychology*, 98(3), 454–468. <http://dx.doi.org/10.1037/a0031909>
- Hofmann, D. A. (2008). Issues in multilevel research: Theory development, measurement, and analysis. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 247–274). Blackwell. <https://doi.org/10.1002/9780470756669.ch12>
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14(4), 1523–1543. <https://doi.org/10.1214/aos/1176350174>
- Homan, A. C., Hollenbeck, J. R., Humphrey, S. E., Van Knippenberg, D., Ilgen, D. R., & Van Kleef, G. A. (2008). Facing differences with an open mind: Openness to experience, salience of intragroup differences, and performance of diverse work groups. *Academy of Management Journal*, 51(6), 1204–1222. <https://doi.org/10.5465/AMJ.2008.35732995>
- Humphrey, S. E., Morgeson, F. P., & Mannor, M. J. (2009). Developing a theory of the strategic core of teams: A role composition model of team performance. *Journal of Applied Psychology*, 94(1), 48–61. <https://doi.org/10.1037/a0012997>

- Ilgen, D. R. (1999). Teams embedded in organizations: Some implications. *American Psychologist*, *54*(2), 129–139. <https://doi.org/10.1037/0003-066X.54.2.129>
- Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, *56*, 517–543. <https://doi.org/10.1146/annurev.psych.56.091103.070250>
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85–98. <https://doi.org/10.1037/0021-9010.69.1.85>
- James, L. R., Demaree, R. G., & Wolf, G. (1993). Rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*(2), 306–309. <https://doi.org/10.1037/0021-9010.78.2.306>
- Jehn, K. A., Rispens, S., & Thatcher, S. M. B. (2010). The Effects of Conflict Asymmetry on Work Group and Individual Outcomes. *Academy of Management Journal*, *53*(3), 596–616. <http://dx.doi.org/10.5465/amj.2010.51468978>
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. <https://doi.org/10.1007/BF02291393>
- Jöreskog, K. G., & Sörbom, D. (1993). *Structural equation modeling with the Simplis command language*. Scientific Software International, Inc.
- Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics*, *25*(3), 1327–1343. <https://doi.org/10.1214/aos/1069362751>
- Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*(1), 126–137. <https://doi.org/10.1037/0022-3514.83.1.126>
- 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*(1), 3–16. <https://doi.org/10.1037/0021-9010.86.1.3>
- Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, *19*(2), 195–229. <https://doi.org/10.5465/amr.1994.9410210745>
- Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*(3), 211–236. <https://doi.org/10.1177/109442810033001>
- Koopmann, J., Lanaj, K., Wang, M., Zhou, L., & Shi, J. (2016). Nonlinear effects of team tenure on team psychological safety climate and climate strength: Implications for average team member performance. *Journal of Applied Psychology*, *101*(7), 940–957. <https://doi.org/10.1037/apl0000097>
- Krijnen, W. P. (2004). Positive loadings and factor correlations from positive covariance matrices. *Psychometrika*, *69*(4), 655–660. <https://doi.org/10.1007/BF02289861>
- Kukenberger, M. R., & D’Innocenzo, L. (2020). The building blocks of shared leadership: The interactive effects of diversity types, team climate, and time. *Personnel Psychology*, *73*(1), 125–150. <https://doi.org/10.1111/peps.12318>
- Larson, N. L., McLarnon, M. J. W., & O’Neill, T. A. (2020). Challenging the “Static” Quo: Trajectories of engagement in team processes toward a deadline. *Journal of Applied Psychology*, *105*(10), 1145–1163. <https://doi.org/10.1037/apl0000479>
- LeBreton, J. M., Moeller, A. N., & Wittmer, J. L. (2023). Data aggregation in multilevel research: Best practice recommendations and tools for moving forward. *Journal of Business and Psychology*, *38*(2), 239–258. <https://doi.org/10.1007/s10869-022-09853-9>
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. <https://doi.org/10.1177/1094428106296642>
- LePine, J. A. (2003). Team adaptation and post-change performance: Effects of team composition in terms of members’ cognitive ability and personality. *Journal of Applied Psychology*, *88*(1), 27–39. <https://doi.org/10.1037/0021-9010.88.1.27>

- LePine, J. A., Hollenbeck, J. R., Ilgen, D. R., & Hedlund, J. (1997). Effects of individual differences on the performance of hierarchical decision-making teams: Much more than g. *Journal of Applied Psychology, 82*(5), 803–811. <https://doi.org/10.1037/0021-9010.82.5.803>
- Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology, 41*(1), 585–634. <https://doi.org/10.1146/annurev.ps.41.020190.003101>
- Li, Y., Li, N., Li, C., & Li, J. (2020). The boon and bane of creative “stars”: A social network exploration of how and when team creativity is (and is not) driven by a star teammate. *Academy of Management Journal, 63*(2), 613–635. <https://doi.org/10.5465/amj.2018.0283>
- Lim, B., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior, 27*(4), 403–418. <https://doi.org/10.1002/job.387>
- Lorinkova, N. M., & Bartol, K. M. (2020). Shared leadership development and team performance: A new look at the dynamics of shared leadership. *Personnel Psychology, 74*(1), 77–107. <https://doi.org/10.1111/peps.12409>
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. <https://doi.org/10.1037/a0012869>
- MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305–325. [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)
- Mao, J., Chang, S., Gong, Y., & Xie, J. L. (2021). Team job-related anxiety and creativity: Investigating team-level and cross-level moderated curvilinear relationships. *Journal of Organizational Behavior, 42*(1), 34–47. <https://doi.org/10.1002/job.2489>
- Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review, 26*(3), 356–376. <https://doi.org/10.5465/AMR.2001.4845785>
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. <https://doi.org/10.1080/00273170903333665>
- Martin, R., Thomas, G., Legood, A., & Dello Russo, S. (2018). Leader–member exchange (LMX) differentiation and work outcomes: Conceptual clarification and critical review. *Journal of Organizational Behavior, 39*(2), 151–168. <https://doi.org/10.1002/job.2202>
- Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*(2), 273–283. <https://doi.org/10.1037/0021-9010.85.2.273>
- Mathieu, J. E., Hollenbeck, J. R., Knippenberg, D. V., & Ilgen, D. R. (2017). A century of work teams in the journal of applied psychology. *Journal of Applied Psychology, 102*(3), 452–467. <https://doi.org/10.1037/apl0000128>
- Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997–2007: A review of recent advancements and a glimpse into the future. *Journal of Management, 34*(3), 410–476. <https://doi.org/10.1177/0149206308316061>
- Moreland, R. L., & Levine, J. M. (1982). Socialization in small groups: Temporal changes in individual-group relations. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 15, pp. 137–192). Academic Press.
- Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th ed.). Muthén & Muthén.
- Neuman, G. A., Wagner, S. H., & Christiansen, N. D. (1999). The relationship between work-team personality composition and the job performance of teams. *Group and Organization Management, 24*(1), 28–45. <https://doi.org/10.1177/1059601199241003>
- Nijstad, B. A. (2009). *Group performance*. Oxford University Press. <https://doi.org/10.4324/9780203872901>
- Peeters, M. A. G., Van Tuijl, H. F. J. M., Rutte, C. G., & Reymen, I. M. M. J. (2006). Personality and team performance: A meta-analysis. *European*

- Journal of Personality*, 20(5), 377–396. <https://doi.org/10.1002/per.588>
- Preacher, K. J., Zhang, Z., & Zyphu, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel sem. *Structural Equation Modeling*, 18(2), 161–182. <https://doi.org/10.1080/10705511.2011.557329>
- Rapp, T., Maynard, T., Domingo, M., & Klock, E. (2021). Team emergent states: What has emerged in the literature over 20 years. *Small Group Research*, 52(1), 68–102. <https://doi.org/10.1177/1046496420956715>
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357. <https://doi.org/10.2307/2087176>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, <https://doi.org/10.18637/jss.v048.i02>
- Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, 7, 1–37.
- Schneider, B., Salvaggio, A. N., & Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology*, 87(2), 220–229. <https://doi.org/10.1037/0021-9010.87.2.220>
- Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications.
- Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *Academy of Management Journal*, 44(2), 316–325. <https://doi.org/10.2307/3069458>
- StataCorp. (2023). *Stata 18 Base Reference Manual*. Stata Press.
- Steiner, I. D. (1972). *Group process and productivity*. Academic Press.
- Thorndike, E. L. (1939). On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. *The American Journal of Psychology*, 52, 122–124. <https://doi.org/https://doi.org/10.2307/1416673>
- Van Kleef, G. A., Homan, A. C., Beersma, B., Van Knippenberg, D., Van Knippenberg, B., & Damen, F. (2009). Searing sentiment or cold calculation? The effects of leader emotional displays on team performance depend on follower epistemic motivation. *Academy of Management Journal*, 52, 562–580. <https://doi.org/10.5465/AMJ.2009.41331253>
- Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, 12(2), 368–392. <https://doi.org/10.1177/1094428107309322>
- Waller, M. J., Okhuysen, G. A., & Saghaian, M. (2016). Conceptualizing emergent states: A strategy to advance the study of group dynamics. *Academy of Management Annals*, 10(1), 561–598. <https://doi.org/10.1080/19416520.2016.1120958>
- Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015). Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices. *Organizational Research Methods*, 18(4), 704–737. <https://doi.org/10.1177/1094428115582090>
- Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11, 1–17. <https://doi.org/10.18637/jss.v011.i10>

Appendix A: Using R to estimate the conceptual mean

```
#' This is a worked example implementing the
# three recommended ways to conduct
# analyses with estimates of the conceptual
# mean using R. We use a dataset from
# Snijders & Bosker (1999) on school language
# test scores of 2287 pupils.
#'
#' We suppose that a researcher has a theoretical
# interest in explaining
# *class-level* language test performance
# as the dependent variable, which
```

```

# requires aggregation of individual-level
test performance (variable `lang`)
# to the class level. We assume here that the
class-level test scores are a reflective team
# construct, which is theoretically problem-
atic, and serves only as an illustration.
#
# We use two predictors that are both fixed
class properties:
# * `GB` = class size
# * `COMB` = whether pupils were taught
in a multi-grade class
# Note: this is not a full and proper
data analysis, and should not be used for
# inference. We skip important steps such as
assumption checks. It merely
# serves to illustrate the implementation of
the three approaches mentioned in
# the source article.
#
# First we prepare the environment and
load the data:
if (!require('pacman')) install.packages(
'pacman') # check for presence of package
manager
pacman::p_load(pacman)
p_load(MASS, lme4, tibble, tidyr, dplyr, per-
formance) # dataset, data wrangling, and
preparation
p_load(sandwich, lmer, lavaan) #
packages needed to implement the three
methods
p_load(stargazer) # output
# Load data from MASS package, documen-
tation: ?MASS::nlschools
data(nlschools)
# We then calculate ICC(1) to check for suf-
ficient agreement and save the
# fit object to re-use it in the two-step
method below:
icc_fit_ml <- lmer(lang ~ 1 + (1 | class),
data = nlschools))
# # Method 1: Arithmetic Mean with
Heteroscedasticity Correction
df_grouplevel <- group_by(nlschools, class,
GS, COMB) %>%
summarise(class_lang = mean(lang)) %>%

```

```

ungroup()
# Fit the linear model:
fit <- lm(class_lang ~ GS + COMB, data =
df_grouplevel)
# Implement the HC3 correction to the
standard errors using vcovHC
# and re-calculate t and p using coeftest:
stats_corrected <- coeftest(fit, vcovHC(fit,
type = 'HC3'))
# Side-by-side comparison shows that the
HC3 correction does not impact the
# coefficients themselves. However, here,
the correction does increase standard
# errors although this does not lead to a dif-
ferent conclusion.
stargazer(fit,
stats_corrected,
type = 'text')
# # Method 2: Two-step Method
#
# We re-use the ICC model that we have fit
above, and extract the random
# intercepts using ranef():
# extract the intercepts:
random_intercepts <- ranef(fit_ml)[[1]]
# rename them:
names(random_intercepts) <- 'class_lang.
eb' # eb = empirical bayes
# because the random intercepts only
capture deviation from the overall
# intercept, we add the overall intercept to
make the outcome variables
# comparable between methods:
random_intercepts$class_lang.eb <- rando-
m_intercepts$class_lang.eb + fixef(fit_ml)[1]
# attach to the group level dataset:
df_grouplevel2 <- left_join(df_grouplevel,
rownames_to_column(random_intercepts,
'class'), by = 'class')
# and fit the model:
fit2 <- lm(class_lang.eb ~ GS + COMB,
data = df_grouplevel2)
# Side-by-side comparison with the base
model shows that the two-step methods
# adjusts both the magnitude and standard
error of the coefficients.
stargazer(fit,

```

```

fit2,
type='text')
#' # Method 3: Direct estimation using
Multilevel SEM
#'
#' We use the lavaan package for R, which
allows us to fit models at two levels simultaneously:
model <- "
level: pupil
lang ~ lang # individual-level variance is
discarded
level: class
clang =~ lang # latent group-level variable
impacts individual-level scores, ...
clang ~ GS + COMB # ... and is predicted by
GS and COMB
"
fit <- sem(model, data = nlschools,
cluster='class')
#' The Level-2 regression output is again
similar to the results obtained above.
summary(fit)
#' # Session information:
sessionInfo()

```

Bernard A. Nijstad is professor of organizational behavior at the University of Groningen, the Netherlands. He is interested in (individual and group/team) creativity and innovation, decision making, and team dynamics. Nijstad published on these topics in journals in social and applied psychology and organizational behavior, and has served as associate editor of *Organizational Behavior and Human Decision Processes*.

Astrid C. Homan is a professor and chair of work and organizational psychology at the University of Amsterdam, the Netherlands. Her research interests are diversity, leadership, team processes and outcomes, and deviance. She aims to understand how to effectively manage and stimulate diversity and being different in work settings. She is Associate Editor at the *Journal of Applied Psychology*.

Marc W. Heerdink is an assistant professor of social psychology at the University of Amsterdam, the Netherlands. His research interests are group processes, social influence and emotions. He aims to understand the role of affect in processes of convergence and divergence between and within groups.

Gerben A. van Kleef is a professor of social psychology at the University of Amsterdam, the Netherlands. His main research programs revolve around emotion, power/hierarchy, social norms, conflict, and cooperation. In studying these topics, he combines social-psychological approaches with insights from various other disciplines, including organizational behavior, evolutionary science, biology, behavioral economics, and law. Van Kleef has served as associate editor of *Social Psychological and Personality Science*, *Cognition and Emotion*, and *Organizational Behavior and Human Decision Processes*.