Journal of Management
Vol. XX No. X, Month XXXX 1–28
DOI: 10.1177/0149206315609402

Implications of Observability for the Theory and Measurement of Emergent Team Phenomena

Nathan T. Carter
Dorothy R. Carter
University of Georgia

Leslie A. DeChurch
Georgia Institute of Technology

Many of the most pivotal mechanisms of team success are emergent phenomena—constructs with conceptual origins at the individual level that coalesce over time through members' interactions to characterize a team as a whole. Typically, empirical research on teams represents emergent mechanisms as the aggregate of members' self-report perceptions of the team. This dominant approach assumes members have developed a perception of the emergent property and are able to respond accurately to survey items. Yet emergent phenomena require sufficient time and team interaction before coalescing as perceptible team properties. Attempting to measure an emergent property before it is perceptible can result in inaccurate assessments and substantive conclusions. Therefore, a key purpose of this study is to develop a better understanding of the underlying characteristics of emergent team phenomena that give rise to their emergence as perceptible and, thus, accurately measurable team characteristics. We advance a conceptual framework that classifies emergent team properties on the basis of the degree to

Acknowledgments: The authors would like to thank the editor and two anonymous reviewers for their patient, thoughtful, and transformative comments on earlier versions of this manuscript. The work was dramatically improved by their insights. Work on this research for L. A. DeChurch was supported in part by the Army Research Institute for the Social and Behavioral Sciences under contracts W5J9CQ-12-C-0017 and W5J9CQ-12-0002. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documents. This material is based upon work conducted by L. A. DeChurch and D. R. Carter, which is supported by the National Science Foundation under Grants SES-1219469 and SBE-1063901. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Supplemental material for this article is available at http://jom.sagepub.com/supplemental

Corresponding author: Nathan T. Carter, Department of Psychology, University of Georgia, 323 Psychology Building, Athens, GA, 30605, USA.

E-mail: ntcarter@uga.edu

which the construct manifests in overtly observable behaviors, positing that more observable emergent team phenomena require less interaction before emerging as ratable team properties compared to constructs that are less easily observed. Leveraging advances in measurement modeling, we test our conceptual framework in a laboratory sample and a quasi–field study sample, demonstrating a multilevel measurement approach that evaluates the emergence of shared team properties across measurement occasions. Results suggest the observability of emergent team properties is a crucial determinant of the relative speed at which constructs emerge as recognizable, ratable properties of the team.

Keywords: *teams; emergence; processes; emergent states; item response theory*

The increasing emphasis on teamwork in organizations has prompted substantial research attention toward uncovering the characteristics of teams that predict team success (Humphrey & Aime, 2014; Mathieu, Maynard, Rapp, & Gilson, 2008). In particular, much prior research on teams has emphasized the importance to team effectiveness of emergent shared team properties, such as members' relatively homologous perceptions of their team's affect, teamwork behaviors, or collective cognition. These are constructs with theoretical origins residing at lower (e.g., individual) levels of analysis that arise over time, through compositional aggregation processes, to characterize the team as a whole (Kozlowski & Klein, 2000). Most empirical studies involving emergent shared team properties have operationalized these properties using the aggregate (e.g., mean) of members' self-report perceptions of their team, justifying aggregation by demonstrating a certain level of agreement across raters (e.g., using interrater agreement indices).

Underpinning this widespread measurement approach is the critical assumption that team members have developed a perception of the emergent property and, thus, are able to respond accurately to perceptual measures. However, we argue that this may not always be a valid assumption, particularly in newly formed teams. Indeed, theories of emergence in organizations emphasize that emergent phenomena in teams require sufficient time and team interaction before coalescing as team-level properties (Arthur, Bell, & Edwards, 2007; Chiocchio & Essiembre, 2009; Cronin, Weingart, & Todorova 2011; Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013; Morgeson & Hofmann, 1999). Given this feature of emergent team properties, it follows that there are periods of time (e.g., initial stages of team development) in which members are not yet capable of providing accurate ratings of these constructs. Simply put, members cannot respond accurately to self-report measures of an emergent property without observing or experiencing an adequate record of relevant interaction.

Moreover, using self-report measures to assess an emergent team property prior to its emergence can result in inaccurate representations of the construct and potentially inaccurate substantive conclusions. Yet although it is common practice to evaluate the homogeneity of team members' ratings of emergent team properties, the assumption that members are capable of providing accurate ratings of these constructs is not evaluated in extant empirical research. This lack of attention to rater accuracy may stem from poor theoretical specificity with regard to when (e.g., in team development) measurement of certain types of emergent team properties might be appropriate, as well as an absence of methodological approaches

with which to empirically evaluate the accuracy of members' ratings. This study aims to address both of these concerns.

First, we advance a guiding conceptual framework that classifies emergent team properties on the basis of the degree to which the construct manifests in overtly *observable* behaviors. We posit that those emergent team phenomena that prior theory has classified as relatively *more* observable (i.e., teamwork processes; Marks, Mathieu, & Zaccaro, 2001) will require relatively *less* team interaction experience before emerging as recognizable and, thus, ratable team properties. This framework offers guidance for researchers wishing to specify a priori the degree to which teams will be capable of accurately evaluating emergent team properties. Second, we demonstrate the use of a *multilevel measurement approach* that evaluates the emergence of shared team properties with regard to *both* the degree to which respondents accurately rate the construct and the homogeneity of members' ratings. In two studies, we use this methodological approach to evaluate our theoretical framework by assessing team members' rating accuracy for emergent shared team properties measured across time.

Theory and Hypothesis Development

In team contexts, members' actions meet in space and time, resulting in discrete interpersonal interaction events (Allport, 1967; Morgeson & Hofmann, 1999). It is through these interaction events that emergent team properties—hypothetical conceptions of phenomena whose existence is inferred on the basis of observable features or actions of the team (Ghiselli, 1964; MacCorquodale & Meehl, 1948; Morgeson & Hofmann)—begin to coalesce as perceptible characteristics of the team as a whole. For team members to qualify as accurate raters of emergent team properties, they must have observed sufficient evidence relevant to the construct (Arthur et al., 2007; Campbell, 1955; Kozlowski & Klein, 2000). Indeed, philosophical accounts of emergence emphasize that one of the defining characteristics of emergent constructs is that they "are recognized by showing themselves, i.e., they are ostensively recognized" (Goldstein, 1999: 50). In the following, we delineate the potentially negative implications of attempting to represent emergent team properties using members' self-report survey responses before the properties are recognizable. Then, we develop a theoretical framework for conceptualizing the relative speed with which emergent phenomena in teams are likely to emerge as perceptible and, therefore, measureable team properties.

Implications of Measuring As-Yet Imperceptible Emergent Team Properties

If team members have not experienced sufficient interaction, as might be the case in initial phases of team development, they may find it "difficult to accurately gauge" (Ellis, 2006: 578) emergent properties of their team. This inaccuracy in rating can result in measurement error and potentially inaccurate substantive conclusions. Indeed, prior research showing stronger effects on team outcomes (D'Innocenzo, Mathieu, & Kukenberger, in press; Kanawattanachai & Yoo, 2007) and higher levels of interrater agreement (Arthur et al., 2007) for emergent team properties measured *later* as opposed to *earlier* in team development is consistent with the idea that measurement error is more prevalent early on.

A parallel can be drawn between asking team members to rate team constructs that are not yet perceptible and the concept of *untraitedness* in personality theory (Baumeister & Tice,

1988). Untraitedness represents the extent to which a particular trait is, or is not, possessed by an individual. Theoretically, individuals who are untraited (i.e., do not possess the trait) would show unstable ratings of themselves on trait-relevant survey items because their behavior is more strongly governed by other factors. Attempting to assess the level of a trait for an untraited individual results in an aberrant pattern of responses that would be poorly predicted by measurement models (see Drasgow, Levine, & Williams, 1985) and cannot be used to accurately place the individual on the personality scale (Drasgow, Levine, & Zickar, 1996; Reise & Waller, 1993). Similarly, team members' ratings of an emergent team property that has yet to be perceived are likely to yield aberrant response patterns that do not accurately reflect the teams' standing (i.e., level) on the construct. Aberrant response patterns can result in measurement error. As explained in the following, measurement error can limit the meaningfulness of team-level scores, the utility of other indicators of emergence for shared team properties (e.g., interrater agreement indices), and the validity of substantive conclusions concerning team functioning.

Implications for Shared Emergent Team Properties

Much prior research has represented collective constructs in models of team effectiveness using the aggregate of members' responses to a self-report perceptual measure of the construct. Inherently, this approach implies a conceptualization of the construct as a *shared* emergent team property—a team characteristic that emerges through relatively simple *compositional* aggregation processes (Kozlowski & Klein, 2000). Shared emergent team properties, in contrast to configural (i.e., patterned) properties, are isomorphic across levels of analysis, such that both lower- and higher-level manifestations share a common meaning and nomological network and the structure of the phenomena is relatively homologous across lower-level units (Kozlowski & Klein).

Empirical studies of teams that represent emergent shared team properties using the aggregate of members' self-reported perceptions of the property often claim evidence of emergence, and therein justification for aggregation, by establishing the homogeneity of members' responses using interrater agreement indices, such as the intraclass correlation (ICC; LeBreton & Senter, 2008). Correctly, James argues that "the use of aggregates . . . is predicated on demonstrating perceptual agreement because agreement implies a shared assignment of psychological meaning" (1982: 228). The ICC assesses the degree to which members' self-report ratings are more strongly a function of the team than a function of individual differences, providing necessary evidence that a measured construct reflects an integrated whole with a homologous structure. However, interrater agreement indices do *not* empirically verify that team members have developed an accurate perception of an emergent team property.

Measures of agreement, like many other statistical approaches in the social sciences (e.g., regression) are based on the assumption that members' responses are error-free indicators of the construct of interest—an unlikely assumption for emergent team properties measured *before* they are perceptible. Violating the assumption that measures of an emergent team property are error free can yield inaccurate evaluations of sharedness using interrater agreement indices in at least two ways. First, although ICCs are robust to measurement errors that are *random* between teams (Fox, 2008), measurement error that varies *systematically* between

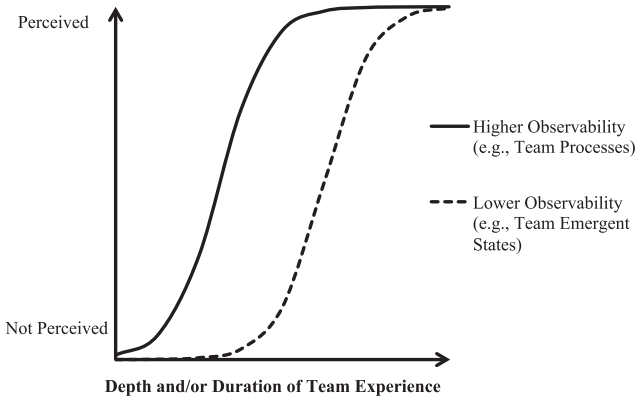
teams is confounded with true differences between teams; measurement error that varies between *individual* team members is confounded with true differences in team member perceptions (i.e., at the individual level). Confounding construct-relevant between-team variance with systematic between-team error variance can result in *overestimation* of interrater agreement. In fact, any between-team variance, whether erroneous or accurate, will be included in between-team variance estimates and to the extent that such erroneous variability exists, will artificially inflate the ICC. Although systematic between-team measurement error is unlikely when members are capable of providing accurate assessments of a team construct, if members have not yet experienced sufficient team interaction to provide accurate ratings, it is likely that sources of systematic error that vary by team would be reflected in interrater agreement indices. For example, when rating interpersonal team trust, raters who have not experienced sufficient team interaction might rely on factors that do not reflect trust per se, such as whether their fellow team members appear likeable or whether they experienced a positive mood during initial interactions. These factors might differ between teams, generating between-team variance in the measured variable; however, the survey responses do not necessarily reflect an informed perception of team trust. As such, reaching current cutoffs for justifying aggregation (e.g., ICC greater than .10; LeBreton & Senter, 2008) is possible as a result of systematic between-team measurement error alone.

Second, confounding true differences in team members' perceptions of their teams with individual-level errors of perception *within* teams can result in *underestimation* of agreement. For instance, if teams have not experienced sufficient interaction to fully determine how well their team shares information, some of the variability across members' ratings of team information sharing would be erroneous and not reflective of members' true perceptions. However, some portion of this variance *will* reflect the raters' true perception of the construct. Confounding these two sources of variance (true vs. error) results in lower than actual estimates of agreement. A more technical explanation of the problem of estimating interrater agreement, given systematic between-team measurement error or individual-level measurement error, can be found in the online supplemental appendix. In summary, attempting to measure an as-yet imperceptible shared emergent team property can result in inaccurate assessments of agreement and, more critically, inaccurate substantive conclusions. Thus, it is critical to specify the likelihood that focal constructs in theoretical models will have emerged *prior to measuring* these constructs using perceptual self-report measures.

Conceptualizing Emergent Team Properties on the Basis of Observability

Theories of emergence suggest that the *speed* with which emergent team properties arise can *vary* on the basis of the nature of the phenomena—different phenomena emerge at different paces (Kozlowski et al., 2013). Thus, we argue that in order to optimally guide measurement decisions, theoretical models of emergence should specify the speed with which focal constructs are likely to emerge. Leading theories of team constructs provide a starting point for this specification. In particular, a significant theoretical advancement in conceptualizing different emergent team properties is the suggestion offered by Marks and her colleagues (2001) that there is a distinction between *team processes* and *team emergent states*. We carry this distinction forward to suggest there is a difference between these types of emergent team properties with regard to the relative speed with which they will be *recognizable*.

Figure 1
Theoretical Relationship Between the Observability of Emergent Team Phenomena and Their Emergence as Ostensively Recognizable (Perceived) Team Properties



Team processes, such as providing backup, coordinating team activities, or sharing information, are depicted in extant literature as members' *observable* behavioral interactions with one another and the task environment (LePine, Piccolo, Jackson, Mathieu, & Saul, 2008; Marks et al., 2001; Mathieu et al., 2008; Tesluk & Mathieu, 1999). Emergent states (e.g., team trust, cognition, cohesion) are depicted as cognitive, affective, or motivational psychological properties of teams that originate in the minds of team members, developing and coalescing as team-level properties while members gain experience interacting with one another (Cronin et al., 2011; Ilgen, Hollenbeck, Johnson, & Jundt, 2005; Marks et al.). In other words, Marks and colleagues' distinction implies that team processes differ from emergent states in that evidence of team process constructs is more readily *observable*, whereas emergent states reflect more *latent* (unobservable) psychological phenomena.

Given that team processes manifest as overtly observable behavioral interactions, these constructs should be rapidly *perceptible* with only a "thin slice" (Ambady, Bernieri, & Richeson, 2010) of relevant team behavior. For example, as soon as team members have engaged in team backup behaviors, these behaviors are immediately recognizable—at least to those who engaged in or were the recipients of the behaviors. In contrast, emergent psychological states are much more idiosyncratic and difficult to identify explicitly. Emergence of latent psychological constructs requires members to draw additional inferences from observed behaviors. For example, developing a shared perception of the degree to which a set of team members are competent and capable of fulfilling different aspects of a team task may involve an extensive amount of time and observation of other members' behaviors in response to varied circumstances. It is feasible that a collective sense of team competence might emerge rapidly. For instance, if members' competence is tested during a sudden and challenging event, a clear sense of team competency may develop immediately. However, in comparison to team processes, latent collective psychological phenomena should generally require greater depth and duration of team interaction and greater inference into the meaning behind behavioral observations before they emerge as perceptible team properties. The solid and dashed lines in Figure 1 depict these two categories of constructs. The solid line

represents relatively more observable emergent team properties, such as team processes. The dashed line represents less observable constructs, such as team emergent psychological states. As depicted in this figure, emergent psychological states require a greater duration of interaction and/or depth of experience (*x*-axis) before they are perceived (*y*-axis) as compared to observable team processes.

The degree to which emergent team phenomena require interaction, observation, and inference before emerging as perceptible—the level of construct observability—should be reflected in members' rating accuracy of these constructs over time. More observable emergent team properties should be rapidly recognizable and, thus, accurately assessed in early as well as later stages of team development. In other words, members' ratings of team processes will be *similarly accurate* across initial and later measurement occasions. In contrast, teams require additional time and interaction before developing accurate perceptions of less observable emergent team properties. As such, members' ratings of emergent psychological states should show *improvement* in accuracy between initial and later measurement occasions aligned with this improvement in perception. Stated formally, we expect:

Hypothesis 1: Relative to emergent team properties that manifest in overtly observable behaviors (e.g., team processes), less observable emergent team properties (e.g., team emergent states) require greater duration and/or depth of team interaction experience before emerging as recognizable and, thus, accurately measurable team properties.

Study 1 Method

Laboratory Teams Sample

In Study 1, we tested our hypothesis using survey responses from a sample of undergraduate student participants in a laboratory experiment conducted at a university in the southeastern United States (N = 648 individuals, 216 teams). Participants were randomly assigned to three-person teams. One member of each team was assigned the role of "team leader" (n = 216 leaders), and the other 2 participants were assigned to one of two unique support roles (i.e., followers; n = 432 followers). Two 3-member teams participated in each experimental session.

Team task. The team task was a PC-based virtual military simulation requiring the two teams to jointly enable a convoy carrying humanitarian supplies to travel safely through a combat zone. During the simulation, the teams distributed supplies to citizen areas and neutralized enemies that threatened the safety of the convoy. After participants were assigned a role in the team, they viewed a training presentation that provided a general overview of the simulation. After the training, the teams completed two 40-min missions in the military simulation. During the first 10 min of each mission, the *transition phases* (Marks et al., 2001), teams had the opportunity to jointly plan and set group goals. During the remaining 30 min of each mission, the *action phases*, participants engaged in the simulation.

Measures

The following team constructs were each measured twice: after the transition or action phase (depending on the dependent variable being measured) in Mission 1 (i.e., Measurement

Occasion 1) and in Mission 2 (i.e., Measurement Occasion 2). For each measure, participants were asked to "think of your teammates when answering the following questions" and to "respond as honestly as possible." Descriptive statistics, scale intercorrelations, and coefficient alphas for all scales are displayed in Table 1. On the basis of our literature review, we classified each construct as either an emergent state or process, each of which is detailed below.

Emergent states: Transactive memory systems credibility and specialization. Transactive memory systems (TMSs), special types of team cognition characterized by accepted divisions of labor among team members for learning, remembering, and communicating relevant knowledge, are thought to benefit team coordination and performance (Hollingshead, 1998; Lewis, 2003). According to Lewis, three dimensions of TMSs can emerge: specialization, coordination, and credibility. *Specialization* involves the structure of differentiated knowledge held by team members, *Credibility* involves team members' confidence in other members' knowledge, and *Coordination* involves the team's tendency and/or ability to work together in a well-coordinated, smooth, and efficient manner (e.g., Lewis; Moreland & Myaskovsky, 2000). After each action phase, participants rated Specialization and Credibility using subscales of Lewis' five-item measure. An example item from the TMS Specialization scale is "Each team member has specialized knowledge of some aspect of our task." An item from the five-item TMS Credibility scale is "I am confident relying on the information that other team members brought to the discussion." Participants responded using a 5-point scale ranging from *strongly disagree* to *strongly agree*.

Emergent state: Team trust. After each action phase, participants rated team interpersonal trust, the willingness of team members to be vulnerable to the actions of other members, using a seven-item measure adapted from the Adams, Thomson, Brown, Sartori, Taylor, and Waldherr (2008) measure of team trust in small military teams. An example item is "My teammates are motivated to protect me." Participants used a 5-point scale ranging from *completely disagree* to *completely agree*.

Team processes: Transition, action, and interpersonal processes. Teams cycle through two recurring episodic phases—transition and action phases—defined by the nature of their associated team interaction processes (Marks et al., 2001): *Transition processes*, interactions typifying transition phases, include interpretation and evaluation of the team's mission, identification and prioritization of team goals, and strategy formulation and planning; *action processes*, typifying action phases, include monitoring aspects of the team's goal progress, systems, and other team members and coordinating interdependent actions. A third category of team processes, *interpersonal processes*, such as conflict and affect management or encouraging motivation and confidence among team members, are relevant throughout both phases. Meta-analytic evidence (LePine et al., 2008) supports the multidimensional hierarchical structure of team processes proposed by Marks et al. and demonstrates positive relationships between team processes and important team outcomes, such as team performance and member satisfaction (LePine et al.). After each transition phase (after the first 10 min of each mission), participants rated the degree to which their team engaged in transition processes; after each action phase (after each 40-min mission ended), participants rated the degree to which their team engaged in action and interpersonal processes. These scales

Table 1

Study 1: Scale Intercorrelations, Coefficient Alphas, Means, and Standard Deviations at Each Measurement Occasion

	Occasion 1							Occasion 2								
Scale	1	2	3	4	5	6	7	1	2	3	4	5	6	7	*M*	*SD*
Occasion 1																
1. TMS Specialization	*.57*														3.86	0.53
2. TMS Credibility	.33	*.72*													3.91	0.61
3. Team Trust	.38	.54	*.87*												3.62	0.68
4. Transition Processes	.18	.24	.29	*.75*											3.84	0.64
5. Action Processes	.33	.47	.60	.28	*.82*										3.42	0.72
6. Interpersonal Processes	.24	.42	.60	.27	.58	*.85*									3.54	0.85
7. Information Sharing	.36	.45	.64	.26	.59	.57	*.85*								3.88	0.77
Occasion 2																
1. TMS Specialization	**.59**							*.73*							4.10	0.57
2. TMS Credibility	.22	**.50**						.45	*.72*						4.02	0.62
3. Team Trust	.34	.35	**.54**					.51	.54	*.89*					3.96	0.64
4. Transition Processes	.16	.16	.24	**.41**				.36	.35	.39	*.87*				3.79	0.72
5. Action Processes	.33	.30	.42	.29	**.42**			.53	.57	.65	.44	*.87*			3.88	0.72
6. Interpersonal Processes	.24	.32	.39	.31	.34	**.54**		.40	.41	.60	.41	.64	*.86*		3.92	0.81
7. Information Sharing	.31	.33	.39	.18	.33	.37	**.41**	.48	.51	.62	.39	.60	.54	*.86*	4.17	0.69

Note: Italicized values are scale coefficient alphas; values in boldface are test-retest correlations. All correlations are significant (at the $p < .01$ level). TMS = transactive memory system.

were developed by Mathieu and Marks (2006) to correspond to definitions provided in the Marks et al. taxonomy. Participants responded to the prompt "To what extent does our team actively work to do the following" using a 5-point scale ranging from *not at all* to *to a very great extent*. Example items from the three-item Transition Processes scale, the four-item Action Processes scale, and the three-item Interpersonal Processes scale are "To what extent does our team actively work to identify the key challenges that we expect to face?", "To what extent does our team actively work to assist each other when help is needed?", and "To what extent does our team work to actively encourage each other to perform our very best?", respectively.

Team process: Information sharing. Meta-analytic evidence shows that the team process of Information Sharing positively predicts team performance, cohesion, decision satisfaction, and knowledge integration (Mesmer-Magnus & DeChurch, 2009). In Study 1, after each action phase, participants used Bunderson and Sutcliffe's (2002) three-item measure to evaluate the degree to which their team shared information. An example scale item is "Information used to make key decisions was freely shared among members of the team." Participants responded using a 5-point scale ranging from *very strongly disagree* to *very strongly agree*.

Analytic Conceptual Framework

Our hypothesis argues that more observable emergent team properties (e.g., team processes) require less duration of interaction experience before emerging as recognizable and accurately measurable as compared to less observable emergent team properties (e.g., emergent states). This hypothesis implies that teams will be able to accurately respond to self-report measures of team processes relatively more quickly than to measures of emergent states. To evaluate this ordinal distinction, we harness psychometric theory, which is grounded in the psychophysical tradition of determining how well persons are capable of making discriminations regarding perceived phenomena (Jones & Thissen, 2007). Specifically, we draw on latent variable theory to show how gains in rating accuracy between measurement occasions manifest as individual calibration and team calibration effects.

Individual calibration. When measuring team constructs with self-report referent-shift surveys, the target of measurement is the team, and a single member's rating reflects his or her perception of the team's standing with regard to the construct, θ_w, the within-subjects latent factor. When a team member *perceives* an emergent team property, his or her responses to survey items assessing the property will be highly intercorrelated because the responses are the outcome of a common cause (i.e., the perception of the property, θ_w). However, when an emergent team property is not yet perceptible, members' item scores either will be loosely correlated or will be correlated as a result of a *different* common cause (i.e., a construct other than the emergent team property). In terms of latent variable theory, when there is a clearly ratable target, item-level scores will load highly onto θ_w; when there is no clearly ratable target, loadings of item scores onto θ_w will be low. We operationalized the accuracy of team member ratings of their own perceptions, λ_w, as the extent to which items load onto θ_w. Thus, an *individual calibration effect* refers to an increase in the strength of relationship between

members' ratings and their true perceptions of the team between measurement occasions (i.e., an increase in λ_w).

Team calibration. Team contexts are inherently multilevel. Thus, the relationship between individual members' responses and the team-level emergent property, referred to here as θ_T, is relevant to understanding team construct emergence. In multilevel latent variable accounts of measurement, there is not only a loading linking the item response to the individual perception of the team, θ_w, but also a loading, which we refer to as λ_T, that links the item response to the team-level variable, θ_T. We operationalize the accuracy of team member ratings of the team-level construct, λ_T, as the extent to which items load onto θ_T, the between-subjects latent variable. As individual team members become better raters of their own perception of the team, they also become more accurate raters of the team as a whole, leading to a higher loading between their item responses and θ_T. Thus, a *team calibration effect* refers to increases in the strength of relationship between ratings and the team-level construct (i.e., increase in λ_T). In summary, an increase in the relationship between team member ratings with both θ_w and θ_T (i.e., λ_w and λ_T) signifies an improvement in accuracy across measurement occasions.

Hypothesized effects. Our hypothesis suggests team processes will show a different pattern of loadings of item scores onto θ_w and θ_T as compared to emergent states. Measures of team processes should show similarly high loadings of item scores onto θ_w and θ_T during initial and later measurement occasions because the rating target is clear almost immediately. In contrast, measures of psychological emergent states should show significant *increases* in loadings (i.e., *individual and team calibration effects*) between initial and later measurement occasions because the clarity of the rating target improves. Lastly, we expect individual calibration effects will be smaller than team calibration effects because λ_T reflects the convergence of all team members' ratings onto the team-level construct of interest, λ_T, rather than a single persons' perception, λ_w.

Analytic Model and Estimation Approach

To test for individual calibration and team calibration effects between measurement occasions, we utilized multiwave multilevel (MWML) item factor analysis (MWML-IFA; Yang, Monroe, & Cai, 2012), a recent development in latent variable modeling. The MWML-IFA model allows estimation of the individual-level latent variable, θ_w, which estimates each team member's perceptions of his or her team, along with the team-level latent variable, θ_T, which represents an estimate of the emergent property for each team. The distribution of θ_w is assumed to be normal with unit variance (i.e., $\sigma_w = 1$), whereas the standard deviation of the team-level latent variable, σ_T, is freely estimated. In multiwave applications, the mean of the individual-level latent variable (μ) must be fixed to 0 for at least one measurement occasion. We set the mean of θ_w in the last occasion to be 0 to serve as a referent for the mean of θ_w for the previous measurement occasion.

For each item in the survey, the strength of association between the individual-level latent variable and the item response is represented by a Level 1 loading, or the *discrimination* parameter, λ_w. The strength of association between the team-level latent variable and the item

response is represented by a Level 2 loading, or *slope*, λ_T. Finally, for each item, the MWML-IFA estimates *intercept* parameters, τ. Each item has one fewer intercept parameter than the number of available response options. For example, a five-option item would have four τ parameters; each intercept represents the point on the trait continuum at which the probability of responding to one option becomes higher than the option below it. These three item parameters (i.e., λ_w, λ_T, and τ) take account of the idiosyncrasies in the measurement properties of the surveys at each time point, allowing for purified estimates of the individual- and team-level latent variables (i.e., θ_w and θ_T). That is, the item parameters act as weights that statistically control for item-specific measurement error at both levels of analysis.

The MWML-IFA model is similar to a multilevel confirmatory factor analysis (CFA) model for more than one measurement occasion. The key difference between these models is that the MWML-IFA model links latent traits and item responses using an item response theory (IRT) model. This distinction enables us to isolate the individual and team calibration effects that are central to our hypothesis. Specifically, past simulations demonstrate that CFA models tend to confound differences in loadings (λ_w) with differences in intercepts (τ), whereas IRT models are capable of distinguishing between the two (see Kankaras, Vermunt, & Moors, 2011).

The MWML-IFA model also enables calculation of an ICC on the basis of the variance of the purified estimates of the individual- and team-level latent variables, $ICC(\theta) = \sigma^2_T/(\sigma^2_T + \sigma^2_w)$. The form of $ICC(\theta)$ corresponds to the ICC(1). However, whereas ICC(1) utilizes the variance of observed scores, representing an amalgam of true variance and variance driven by measurement error, the $ICC(\theta)$ is calculated from the variances in θ_w and θ_T. These individual- and team-level latent variable estimates isolate the construct-relevant component of the observed score. This approach helps avoid confounding construct-relevant variance with systematic between-team error variance or individual-level error variance. The $ICC(\theta)$ results in a more accurate assessment of team member consensus by isolating the construct-relevant variance in ratings.

The MWML-IFA model is a general latent variable approach that can incorporate any IRT model. We utilized the graded response model (GRM; Samejima, 1969) for estimation, as this model is often used for polytomous self-report surveys (see Zickar, 1998). We estimated item parameters for the MWML-IFA using marginal maximum likelihood estimation. To test for individual and team calibration effects, we conducted analyses consistent with measurement equivalence/invariance (ME/I) analysis using the flexMIRT software program (Cai, 2013). We utilized a hierarchical model comparison approach using the likelihood ratio (LR) test statistic (Stark, Chernyshenko, & Drasgow, 2006), which tests for differences in relevant parameters between measurement occasions. This process is summarized below:

1. Identify one appropriate anchor item for each measure to establish a scale for the latent variables by the method suggested by Meade and Wright (2012) in single-level IRT analyses.
2. Estimate a fully freed baseline model in which all item parameters (with the exception of the parameters for the anchor item) are allowed to vary between measurement occasions.
3. Constrain Level 1 item locations (τ parameters) to be equal between measurement occasions; this step controls for differences in τ parameters, which are less relevant to emergence.
4. Constrain the Level 2 slopes (λ_T parameters) to be equal between occasions and compare to the model in No. 3 via the LR test; if this model fits worse than the model in No. 3, then Level 2 slopes differ significantly between occasions, indicating *team-level calibration* occurred.

Table 2
Study 1: Tests and Effect Sizes for Each Scale Between Occasions

Measure	Test	LR	*df*	*p*	φ	Yes/No
Team Trust	L2 Calibration?	15.26	6	.007	.11	Yes
	L1 Calibration?	24.88	6	< .001	.13	Yes
TMS Credibility	L2 Calibration?	14.58	4	.002	.11	Yes
	L1 Calibration?	22.72	4	.000	.13	Yes
TMS Specialization	L2 Calibration?	19.49	4	< .001	.12	Yes
	L1 Calibration?	14.22	4	.003	.11	Yes
Information Sharing	L2 Calibration?	0.49	2	.391	.02	No
	L1 Calibration?	0.47	2	.395	.02	No
Interpersonal Process	L2 Calibration?	2.60	2	.136	.04	No
	L1 Calibration?	10.14	2	.003	.08	No
Transition Process	L2 Calibration?	8.08	2	.009	.08	No
	L1 Calibration?	16.76	2	< .001	.11	Yes
Action Process	L2 Calibration?	0.51	2	.387	.02	No
	L1 Calibration?	2.82	2	.122	.05	No

Note: LR = likelihood ratio test; L2 calibration = model with item locations and Level 2 slopes fixed to be equal over time; L1 calibration = model with all parameters, including Level 1 discriminations, fixed to be equal over time; yes = the effect was observed (i.e., φ > .10 and $p < .05$); no = the effect was not observed; TMS = transactive memory system.

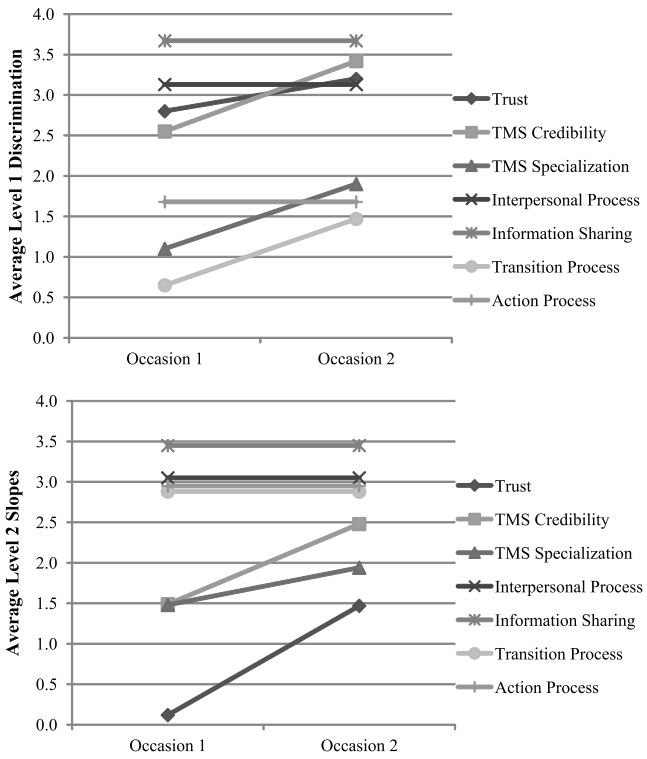
5. Constrain Level 1 discriminations (λ_w parameters) to be equal between occasions; compare to the model in No. 4 via the LR test; if the model fit is worse than the model in No. 4, discriminations differ significantly between occasions, indicating *individual calibration*.
6. Determine the best-fitting model tested in Nos. 2 through 5 and estimate ICC(θ).

Preliminary model-data fit analyses. Because the MWML-IFA model is a recent development, there is, at present, little guidance for evaluating absolute model-data fit. Therefore, prior to conducting our ME/I analyses, we conducted single-level model-data fit analyses using the IRTPRO program (Cai, Thissen, & du Toit, 2011) to ensure that the graded response IRT model showed good fit to the data using the recently developed M_2 fit statistic (Cai & Hansen, 2013), which allows for the calculation of the root mean square error of approximation (RMSEA) global fit statistic in IRT models. RMSEAs showed model-data fit of the GRM was adequate, falling between .05 and .08 for all scales, with the exception of TMS Specialization and Information Sharing, with RMSEA of .11 (90% confidence interval, CI, of .10 to .12) and .12 (90% CI of .11 to .13), respectively, both slightly above the criteria for a mediocre model. Thus, we moved forward with MWML-IFA analyses, which would be expected to have equivalent, or better, model-data fit than the single-level model.

Study 1 Results

Table 2 displays the results of LR statistical significance tests and associated effect sizes (Cramer's φ, the square root of the chi-square-distributed LR statistic divided by *N*) for the

Figure 2
Study 1: Individual and Team Calibration Effects



Note: The top panel shows the average Level 1 discrimination parameter at each measurement occasion (individual calibration); the bottom panel shows the average Level 2 slope at each measurement occasion (team calibration). TMS = transactive memory system.

multilevel IRT model.[1] To control for Type I error rates in LR tests, we considered effects to be significant only if they showed statistical significance (i.e., $p < .05$) and had an effect size, φ, greater than .10.

In alignment with our hypothesis, the measures of Team Trust, TMS Credibility, and TMS Specialization all showed a pattern of results consistent with an increase in accuracy between measurement occasions: (a) individual calibration effects (i.e., increases in λ_w between measurement occasions) and (b) team calibration effects (i.e., increases in λ_T between measurement occasions). Increases in λ_w and λ_T are shown in Figure 2.

Consistent with our hypothesis, results showed that none of the team process measures exhibited the combination of individual and team calibration effects observed in measures of emergent states. Neither effect was observed for interpersonal processes or action processes. The Transition Processes measure showed the individual calibration effect, but not the team calibration effect, which as noted previously is more central to the notion of observability of a team-level property.

Table 3
Study 1: Comparison of Traditional and Item Response Theory–Based Intraclass Correlations

	Occasion 1			Occasion 2		
Construct	ICC(1)	ICC(θ)	ICC(1)-ICC(θ)	ICC(1)	ICC(θ)	ICC(1)-ICC(θ)
Team Trust	.16	.09	.07	.21	.21	.00
TMS Credibility	.01	.50	−.49	.20	.42	−.22
TMS Specialization	.03	.43	−.40	.01	.43	−.42
Information Sharing	.04	.09	−.05	.12	.16	−.04
Interpersonal Processes	.11	.10	.01	.09	.14	−.05
Transition Processes	.09	.17	−.08	.18	.21	−.03
Action Processes	.14	.25	−.11	.12	.27	−.15

Note: ICC(1) is the intraclass correlation based on the observed scores; ICC(θ) is the intraclass correlation based on item response theory scores controlling for measurement error; ICC(1)-ICC(θ) is the difference between the two. TMS = transactive memory system.

Table 3 displays the ICC(1) based on observed scores using maximum likelihood estimation of variance components compared to the ICC(θ) based on the best-fitting of the measurement models tested. For all scales other than Team Trust, the ICC(1) underestimated agreement compared to the ICC(θ). ICC(1) was much lower than ICC(θ) in both measurement occasions for the TMS Credibility and TMS Specialization measures, suggesting that measurement error confounded the estimation of agreement. The underestimation of the ICC based on observed scores is consistent with our expectations regarding the influence of individual-level measurement error on the estimation of ICC(1). The Team Trust measure showed ICC(1) as an overestimate, which is consistent with our expectations for the influence of between-team measurement error on the estimation of ICC(1). This idea is further bolstered by the finding that Team Trust showed by far the largest increase in λ_T, suggesting that between-team error is indeed associated with overestimation of agreement using the ICC(1).

Study 1 Discussion

Results of Study 1 suggest that, in comparison to team processes, constructs commonly classified as psychological team emergent states require more extensive team interaction experience before emerging as recognizable, and *ratable*, team properties. On the basis of prior theory (Cronin et al., 2011; Marks et al. 2001), we argue that the different patterns of calibration effects for team processes versus emergent states stem, in part, from differences in their observability.

In alignment with our hypothesis, team members' ratings of emergent state constructs showed significant individual and team calibration effects between measurement occasions, signifying an improvement in members' ability to accurately rate these team properties after additional interaction. In contrast, none of the team processes showed significant team-level calibration effects. Only one process variable—team transition process—showed a significant individual-level calibration effect. However, this measure was the only scale administered during the first 10 min of each simulation exercise, whereas all other measures were

administered at the end of each 40-min simulation. Thus, the transition process effect may have stemmed from the extremely short duration of interaction members experienced prior to initial evaluation. This effect reflects only greater consistency in individual perception rather than greater accuracy in rating the team-level construct.

A Potential "Continuum" of Observability

Although our results provided general support for our hypothesis, the observed effects did not suggest a clear *dichotomous* distinction between team emergent states and team processes. Team Trust, which involves members' judgments about their own and others' willingness to be vulnerable to others' actions, showed the largest calibration effects. TMS Credibility also had sizable calibration effects. However, the effects for TMS Specialization were much smaller and more similar to effects observed in team process measures. This suggests teams were more rapidly capable of evaluating the level of TMS Specialization than they were TMS Credibility or Team Trust. The differences in these effect sizes suggest that these constructs might exist along a *continuum* of observability ranging from more observable to less observable emergent team phenomena.

Indeed, Mathieu et al. (2008) noted that TMS is a "blended" mediator in models of team effectiveness such that TMS has features associated with both emergent states and processes. TMS Credibility involves members' perceptions regarding the degree to which they can rely on the information provided by other members. TMS Specialization involves a cognitive representation of the degree to which other members possess unique information or, instead, all possess identical information. Measures of Team Trust and TMS Credibility may require members to make additional inferences into the underlying *meaning* of members' behaviors to determine, for example, whether they trust the reliability of other members' contributions. Rating TMS Specialization may require significantly less inference given that TMS Specialization items require raters to evaluate the degree to which they perceive expertise is distributed among team members but not their perception of members' ability to use that expertise competently. Finally, Coordination may be the most similar to a process variable in that it concerns the observable interactions involved in coordinating team tasks. A limitation of Study 1 is that Coordination was not assessed.

Moreover, although teams research sometimes casts emergent team properties as either an emergent psychological state or a team process, our results suggest that these constructs might be better conceptualized along a continuum, ranging from low to high levels of observability. In Study 2, we further evaluate this continuum by assessing the relative improvement in rater accuracy over time for additional measures of shared emergent team properties.

Study 2

In Study 2, we evaluate calibration in self-report measures of Affect-Based Team Trust, Cognitive-Based Team Trust, TMS Credibility, Collective Efficacy, TMS Specialization, and TMS Coordination using a quasi–field study teams sample with three measurement occasions. The inclusion of a greater number of time points and wider variety of emergent state measures allows us to further assess the idea that some constructs emerge more quickly than others by comparing the rate at which individual-level factor loadings, λ_w, and team-level

factor loadings, λ_T, increase over time. To further investigate the idea that emergent team properties exist along a common continuum of observability, we obtained independent subject matter expert (SME) ratings of the observability of the item content for all measures. By correlating SME ratings of observability with changes in the λ_w and λ_T indicators of emergence between measurement occasions, we provide a more direct test of the idea that less observable constructs require more time to emerge than more observable constructs.

Study 2 Method

Quasi–Field Study Teams Sample

We assessed calibration in measures of emergent states using survey responses from a sample of undergraduate- and masters-level student participants in an 8-week quasi–field study conducted across two universities (N = 160 individuals, 76 teams). Students participated in the study as part of their course grade in one of four separate courses: an ecology course at a northeastern university in the United States, two different sections of a social psychology course at the same U.S. university, and a business management course at a university in France.

Team task. The 8-week project commenced as follows. First, students were randomly assigned to a three- or four-member team of fellow classmates. Second, each team was randomly matched with three other teams from the other university courses to form 19 four-team systems. As part of their course grade, each team was assigned the goal of developing their knowledge within their own area of expertise and submitting multiple team deliverables throughout the semester. The ecology teams analyzed the nature of an environmental problem, the two social psychology teams researched ways to apply attitude and behavior change strategies to their stakeholders, and the business teams researched the value network of individuals and organizations that play a role in the environmental problem. Each four-team system had the shared goal of integrating its expertise to create a written action plan for a policy or product that had a strong potential to positively affect an important environmental issue.

Measures

Participants completed measures of team emergent states at three measurement occasions: Week 3 (n = 129), Week 5 (n = 111), and Week 8 (n = 160). Instructions indicated that participants were to consider their "project team" (i.e., students with whom they worked on the group project) as the referent for these measures and respond using a 5-point scale ranging from *strongly disagree* to *strongly agree*. See Table 4 for descriptive statistics and Table 5 for scale intercorrelations and coefficient alphas.

Affect-based and cognitive-based team trust. Participants rated Affect-Based Team Trust using a five-item measure and Cognitive-Based Team Trust using a four-item measure, both of which were adapted from McAllister (1995). An example Affect-Based Team Trust item is "If I shared my problems with this team, I know team members would respond constructively and caringly." An example Cognitive-Based Team Trust item is "This team approaches our project with professionalism and dedication."

Table 4
Study 2: Means and Standard Deviations for Each Measurement Occasion and Subject Matter Expert (SME) Observability Ratings for Each Scale

	Occasion 1		Occasion 2		Occasion 3		SME Observability Ratings (n = 5)	
Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Affect-Based Team Trust	3.81	0.99	3.81	1.13	3.96	1.14	2.12	0.23
Cognitive-Based Team Trust	3.90	0.98	3.97	1.10	4.05	1.14	2.15	0.26
Collective Efficacy	2.99	0.89	3.09	0.88	3.14	0.89	2.00	0.30
TMS Credibility	3.44	0.79	3.53	0.75	3.46	0.78	2.07	0.30
TMS Specialization	4.04	0.88	4.15	0.85	4.11	0.86	2.75	0.26
TMS Coordination	3.89	1.16	3.92	1.27	3.86	1.42	3.07	0.30

Note: TMS = transactive memory system.

TMS specialization, credibility, and coordination. Participants rated TMS using short versions of Lewis' (2003) TMS scales. TMS Specialization and TMS Credibility scales were the same as those used in Study 1. TMS Coordination was assessed using a three-item scale. However, one item ("There is much confusion about how we should accomplish the task") was excluded as a result of low coefficient alphas across occasions (α = .43, .65, and .65, respectively) and problems caused by this item in estimation of the IRT models. The remaining items were "Our team works together in a well-coordinated fashion" and "We accomplish the task smoothly and efficiently."

Collective efficacy. Participants rated the degree to which they felt collectively efficacious with their team using Edmondson's (1999) three-item scale. An example item from this scale is "With focus and effort this team can do anything we set out to accomplish."

Analytic Conceptual Framework, Model, and Estimation Approach

In Study 2, we use the same MWML-IFA model as in Study 1 to determine the pattern of changes in individual-level factor loadings, λ_w, and team-level factor loadings, λ_T, between measurement occasions. Additionally, we assess the extent to which differences in λ_w and λ_T correlate with SME ratings of observability. Study 2 included three measurement occasions. Consistent with Study 1, the final occasion was used as the referent in all analyses by fixing the mean and standard deviation of the Time 3 within-team latent variable to 0 and 1, respectively. The item in each scale with the highest discrimination across measurement occasions in a single-level IRT analysis was used as the anchor item for the scale (Meade & Wright, 2012).

As a result of concerns regarding the stability of estimates obtained using the marginal maximum likelihood approach to estimation (the approach used in Study 1) with small sample sizes, we utilized a Bayesian approach to estimate the MWML-IFA model called the Metropolis-Hastings Robbins-Monroe (MHRM; see Cai, 2010) algorithm. The MHRM is a Monte Carlo–based approximation to the marginal log likelihood for incomplete data (Houts & Cai, 2013), meaning item parameter estimates are obtained by iterative simulations that

Table 5
Study 2: Scale Intercorrelations and Coefficient Alphas at Each Measurement Occasion

	Occasion 1						Occasion 2						Occasion 3					
Scale	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Occasion 1																		
1. Affect-Based Team Trust	*.89*																	
2. Cognitive-Based Team Trust	**.85**	*.89*																
3. Collective Efficacy	**.73**	**.83**	*.86*															
4. TMS Credibility	**.56**	**.59**	**.64**	*.91*														
5. TMS Specialization	**.51**	**.56**	**.59**	**.59**	*.84*													
6. TMS Coordination	**.71**	**.78**	**.74**	**.56**	**.59**	*.92*												
Occasion 2																		
1. Affect-Based Team Trust	(.33**)	.29**	.38**	.44**	.36**	.21	*.89*											
2. Cognitive-Based Team Trust	.27*	(.31**)	.29**	.36**	.35**	.22*	**.80**	*.87*										
3. Collective Efficacy	.16	.23*	(.28**)	.30**	.28**	.09	**.57**	**.73**	*.82*									
4. TMS Credibility	.17	.25*	.35**	(.44**)	.25*	.12	**.57**	**.68**	**.61**	*.89*								
5. TMS Specialization	.29**	.33**	.34**	.40**	(.44**)	.22*	**.46**	**.56**	**.59**	**.68**	*.76*							
6. TMS Coordination	.30**	.36**	.41**	.48**	.33**	(.25*)	**.77**	**.80**	**.68**	**.64**	**.50**	*.87*						
Occasion 3																		
1. Affect-Based Team Trust	(.33**)	.28**	.17	.18	.19*	.24*	(.42**)	.43**	.34**	.24*	.24*	.37**	*.92*					
2. Cognitive-Based Team Trust	.21*	(.32**)	.18	.11	.16	.26**	.36**	(.46**)	.28**	.18	.24*	.41**	**.84**	*.91*				
3. Collective Efficacy	.22*	.28**	(.22*)	.15	.24*	.19	.26**	.38**	(.36**)	.15	.26**	.38**	**.73**	**.80**	*.85*			
4. TMS Credibility	.07	.09	−.02	(.19*)	.07	−.01	.28**	.33**	.20*	(.34**)	.13	.26*	**.64**	**.61**	**.55**	*.86*		
5. TMS Specialization	.09	.13	.09	.06	(.21*)	−.01	.13	.20*	.25*	.24*	(.32**)	.22*	**.59**	**.55**	**.48**	**.53**	*.75*	
6. TMS Coordination	.24*	.29**	.15	.18	.19*	(.22*)	.31**	.41**	.31**	.14	.21*	(.39**)	**.75**	**.84**	**.76**	**.63**	**.52**	*.96*

Note: Italicized values in the main diagonal are scale coefficient alphas; values in parentheses are test-retest correlations. Scale intercorrelations within measurement occasions are shown in bold. Correlations in bold are significant (at $p < .001$). TMS = transactive memory system.

*$p < .05$.

**$p < .01$.

converge when estimates stabilize.[2] Using the model comparison strategy described in Study 1, we determined the best-fitting model for each measure. The best-fitting model was then used to estimate item parameters.

Preliminary model-data fit analyses. We evaluated the absolute model-data fit of the single-level GRM. Fit statistics for the fully freed baseline model of each scale showed good to moderate model fit, with RMSEAs between .048 and .053 for all scales except the Collective Efficacy and TMS Credibility scales, which had RMSEAs just above the cutoff reasonable fit, .082 (90% CI from .072 to .092) and .086 (90% CI from .076 to .096), respectively.[3]

SME Ratings of Observability

To estimate the observability of the constructs measured in Study 2, we had five SMEs, all doctoral students, rate the difficulty of observing the content of each item for each scale. SMEs were instructed to

> indicate how easy to difficult it would be to accurately observe and rate the item if you were in a team that was in the very early stages of team formation (e.g., one of the first days of working on a project with a team). Imagine a team of between 3 and 6 people. Provide your general feeling of how easy it would be to observe and rate the item.

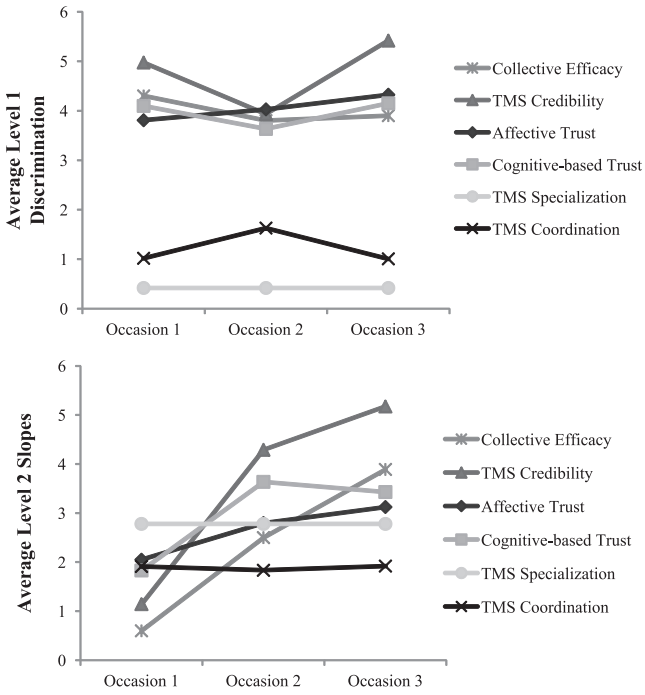
SMEs rated each item on a 4-point scale ranging from *very difficult to observe* to *very easy to observe*. Ratings were highly reliable, with ICC for absolute agreement at .70 and internal consistency of .77. Observability ratings for each scale are shown in Table 4.

Study 2 Results

Table 4 provides a summary of SME ratings of speed of observability for the constructs measured in Study 2. Collective Efficacy, TMS Credibility, Affect-Based Team Trust, and Cognitive-Based Team Trust items showed the lowest SME-rated observability scores, with mean SME ratings of 2.00, 2.07, 2.12, and 2.15, respectively. TMS Specialization and TMS Coordination showed relatively higher ratings of observability, with mean SME ratings of 2.75 and 3.07, respectively. We expected those constructs with relatively lower observability ratings to show larger increases in λ_w and λ_T over time, whereas those constructs with higher observability ratings were expected to show relatively stable λ_w and λ_T estimates over time.

Plots of average λ_T and λ_w parameters are shown in Figure 3. Collective Efficacy showed large and consistent increases in λ_T parameters between each occasion, consistent with its low SME observability score, but no consistent increase in λ_w. TMS Credibility showed a large increase in λ_T parameters between the first two occasions and a much smaller increase between the last two occasions, consistent with its lower observability, but did not show a steady increase in λ_w parameters. Affective trust showed steady, smaller increases in λ_T, consistent with its lower observability score, but negligible increases in λ_w. Cognitive-Based Team Trust showed a large increase in λ_T parameters between the first two occasions but negligible change in λ_T between the last two occasions, whereas plots of the λ_w parameters showed no large changes between occasions. TMS Specialization and TMS Coordination measures showed no increase in λ_T or λ_w, consistent with their higher observability scores.

Figure 3
Study 2: Individual and Team Calibration Effects



Note: The top panel shows the average Level 1 discrimination parameter at each measurement occasion (individual calibration); the bottom panel shows the average Level 2 slope at each measurement occasion (team calibration). TMS = transactive memory system.

The correlation between SME observability ratings and differences in the IRT-based indicator of λ_T (Level 2 calibration) between the first and last occasions was significant, large, and in the expected direction, $r(6) = -.77$, $p = .036$, such that constructs with lower observability ratings showed larger increases in λ_T estimates between occasions. The correlation between observability ratings and differences in λ_w estimates between occasions (Level 1 calibration) was nonsignificant but in the expected direction, $r(6) = -.21$, $p = .348$. Thus, in general, findings for differences in λ_T were consistent with our expectation that relatively less observable constructs would require relatively more depth and duration of team interaction before emerging as recognizable team properties, whereas findings for λ_w were not entirely consistent with our expectation. As noted in the Study 2 Discussion section, this may suggest differences in λ_T estimates between occasions is the more important indicator of emergence.

As in Study 1, we evaluated the difference between ICC(1) and IRT-based ICC(θ). Table 6 shows the ICC(1) and ICC(θ) for each scale at each measurement occasion based on the best-fitting measurement model. For most scales, ICC(1) underestimated agreement compared to ICC(θ). Similar to Study 1, underestimation was largest for TMS Specialization and TMS Credibility. The only notable exception was the Affect-Based Team Trust scale, which

Table 6
Study 2: Comparison of Traditional and Item Response Theory–Based Intraclass Correlations

	Occasion 1			Occasion 2			Occasion 3		
Construct	ICC(1)	ICC(θ)	ICC(1)-ICC(θ)	ICC(1)	ICC(θ)	ICC(1)-ICC(θ)	ICC(1)	ICC(θ)	ICC(1)-ICC(θ)
Affect-Based Team Trust	.15	.01	.14	.26	.04	.22	.09	.28	−.19
Cognitive-Based Team Trust	.02	.12	−.10	.08	.12	−.04	.19	.21	−.02
Collective Efficacy	.00	.06	−.06	.27	.22	.05	.13	.19	−.06
TMS Specialization	.00	.22	−.22	.03	.12	−.09	.00	.11	−.11
TMS Credibility	.00	.07	−.07	.00	.32	−.32	.00	.44	−.44
TMS Coordination	.00	.17	−.17	.31	.41	−.10	.24	.22	.02

Note: ICC(1) is the intraclass correlation based on the observed scores; ICC(θ) is the intraclass correlation based on item response theory scores controlling for measurement error; ICC(1)-ICC(θ) is the difference between the two. TMS = transactive memory system.

overestimated agreement in ICC(1) compared to ICC(θ) for the first two measurement occasions. However, in the final measurement occasion, ICC(1) was lower than ICC(θ). This finding suggests that in earlier measurement occasions, team-level measurement error was large for the Affect-Based Team Trust measure relative to individual-level error, whereas in the later measurement occasion, individual-level error was larger. For all measures except TMS Credibility, the differences between ICC(1) and ICC(θ) decreased or remained near zero over time.

Study 2 Discussion

Study 2 examined the idea that the relative extent of team interaction required for emergent team phenomena to arise as recognizable team properties can be roughly predicted by considering the degree to which teams will experience observable evidence of the construct. Our findings support previous speculation that there is a continuous, rather than dichotomous, distinction between team emergent states and process constructs (e.g., Mathieu et al., 2008).

Results of Study 2 supported the majority of our expectations regarding a continuum of observability. Calibration (i.e., differences in both λ_T and λ_w) for Affect-Based Team Trust was consistent with our expectation: This scale suggested progression in terms of accuracy, with relatively steady increases over time in the relationship between survey scores and the measured variable at both the team and individual level. Results for Collective Efficacy, TMS Credibility, and Cognitive-Based Team Trust were consistent with our expectations at the team level. Team calibration (i.e., differences in λ_T) for these constructs indicated a relatively large increase in accuracy between the first and second measurement occasion and small or negligible changes between the second and third occasion. Additionally, findings of negligible changes in λ_T and λ_w for the more observable TMS Specialization and TMS Coordination constructs were consistent with our expectations. However, in conflict with

Study 1, results did not show evidence of individual calibration for any measures. One potential reason we did not find individual calibration in Study 2 could be the greater lapse of time between team formation and the first measurement in Study 2 compared with Study 1 (3 weeks vs. 40 min, respectively). That is, individual calibration may have occurred prior to the first measurement in Study 2. Individual calibration effects were small in Study 1 and may evince themselves only at *very* nascent stages of team formation.

In sum, we found that observability ratings were highly correlated with λ_T differences between measurement occasions and showed a small correlation with differences in λ_w, suggesting that λ_T is indeed a stronger indicator of emergence. In a follow-up analysis using the same SMEs to conduct ratings for the variables used in Study 1, we found a nearly identical pattern, such that a strong correlation was found between observability ratings and changes in λ_T, $r(7) = -.64$, $p = .061$, and a weak correlation between observability and λ_w, $r(7) = -.07$, $p = .439$.[4]

In addition to further exploring the nature of team emergent states, Study 2 extended upon Study 1 by examining calibration in longer-duration teams with better potential to generalize to real-world organizational populations. Although the sample in Study 1 afforded the many advantages of laboratory studies (e.g., random assignment, increased control, comparable teams), external validity was limited in Study 1 for at least three reasons: (a) the time participants worked together as a team was very short in comparison to real-world teams, (b) the experimental team task had no real-world consequences for members, and (c) measurements of team constructs were taken at only two time points. The quasi–field study setting used in Study 2 had the advantage of providing a stronger motivation for high team performance (i.e., course grades) and a richer set of team experiences and interactions than is possible in the lab. Results of Study 2 suggest construct observability is a realistic concern in quasi–field study and real-world teams samples.

General Discussion

Emergent phenomena are foundational to the study of teams. Scholars have noted that it is vital to accurately specify theoretic conceptions of focal emergent constructs prior to data collection (Kozlowski & Klein, 2000). We maintain that theoretic conceptions should consider *when* focal constructs are likely to be recognizable and, thus, ratable team properties. Our results highlight construct observability as a key characteristic of emergent phenomena that has implications for the speed with which they will emerge as recognizable team properties. Team-level calibration effects were consistent with our hypothesis: During initial stages of team development, teams are relatively more accurate in their ratings of more observable emergent phenomena. Our findings also suggest emergent team properties can be conceptualized along a continuum of observability rather than as discrete categories (i.e., emergent states vs. processes). Finally, we showed that SME observability ratings were strongly related to team calibration effects, suggesting that observability plays a substantial role in the accuracy of members' ratings of emergent team properties.

Theoretical and Methodological Contributions

This study has both theoretical and methodological implications for teams research. First, we improved theoretical understanding of emergent team phenomena by offering construct

observability as a key driver of emergence. Marks and colleagues (2001) alluded to observability as a defining characteristic of emergent team phenomena, classifying team processes as more observable than team emergent states. We extended this conceptualization by directly explicating observability and clarifying its implications for members' ability to perceive and accurately rate emergent team properties over time. Scholars have stressed the need for teams research that better accounts for temporal dynamics, particularly with regard to emergent phenomena (e.g., Kozlowski & Chao, 2012; Mohammed, Tesler, & Hamilton, 2012). By identifying observability as an underlying characteristic of emergent team properties, we provide initial guidance for specifying the temporal aspects of emergent phenomena prior to measurement. Certainly, more research is needed that continues to investigate team temporal dynamics.

A second theoretical contribution is the development of a second conceptual framework that harnesses latent variable theory to specify the complex processes (e.g., individual and team calibration) occurring as shared team properties emerge and how these processes are reflected in the measurement properties of survey instruments. This is an important development given that, although teams research has evolved a sophisticated understanding of the importance of interrater agreement as an indicator of construct emergence (e.g., Chen, Mathieu, & Bliese, 2004; James, 1982; LeBreton & Senter, 2008), insufficient attention has been paid to team members' rating *accuracy*. Although findings in Study 2 were fully consistent only for team calibration, Study 1 showed support for individual calibration effects. In fact, in Study 1, the indices of these effects (λ_w and λ_T changes) were highly correlated, $r(7) = .55$, $p = .058$, as expected given that we theorized that individual accuracy would lead to team accuracy. However, these indices were uncorrelated in Study 2. Individual calibration may not have been detected in Study 2 as a result of longer passages of time between measurement occasions and may appear only at very early stages of team development. Future research should further test and refine this framework to confirm the role of individual calibration in emergence.

Moreover, our results suggest the path to emergence for shared perceptual team properties may involve at least two stages. In the first stage, some degree of perception or comprehension that a theoretical phenomenon exists in the team is formed. As we demonstrated, the ability of teams to recognize a phenomenon will depend on the observability of the construct and the duration and/or depth of experience with the team. In the second stage, members' perceptions of the property need to converge to represent a shared perception. We developed the theoretic concepts of individual and team calibration (i.e., improvement in members' accuracy in team evaluation over time) in order to conceptualize the first stage of this process. As such, this article fills a crucial gap in the conceptualization and understanding of emergence.

Methodologically, this paper demonstrates a novel approach for operationalizing, detecting, and accounting for members' ability to accurately rate shared emergent team properties. The MWML-IFA approach represents a powerful tool for studying emergence in that it is capable of (a) more accurately estimating homogeneity using ICC(θ); (b) estimating team-level scores, θ_T, that are less confounded with measurement error than aggregates of observed scores; and (c) determining whether observed scores can be compared over time (i.e., the more typical use of ME/I analysis). Although the method does not provide a definitive cutoff for determining when a construct has fully emerged, emergence can be inferred from null team calibration effects between measurement occasions and satisfactory ICC(θ).

The MWML-IFA approach shows great promise for understanding and investigating team construct emergence using referent-shift surveys. Although the large sample sizes needed for the IRT-based analyses may be prohibitive to many researchers, there are other latent variable models that can be used to identify and/or account for members' rating accuracy in team-level surveys. Researchers wishing to directly compare observed scores collected at different time points and/or attain more accurate ICC estimates could use MWML CFA to determine whether scores are equivalent between time points and estimate ICCs on the basis of latent variable estimates (as opposed to observed scores). However, past simulations show that CFA-based approaches are not capable of differentiating nonequivalence due to differences in τ parameters from individual calibration effects (Kankaras et al., 2011). Researchers wanting to detect individual calibration could use single-level IRT analyses with smaller samples, but this approach does not enable estimation of team calibration or ICC(θ)s.

Future Research Directions

Of course, in the present study, time is only a *proxy* for team interaction experience, and differences between teams in their experiences, rather than time itself, will be the primary drivers of construct emergence. It is beyond the scope of the current study to identify the exact processes and experiences that enabled team construct emergence—certain experiences may have led some teams to more quickly develop a sense of trust than others, for example. Thus, we echo recent calls (e.g., Hackman, 2012; Kozlowski et al., 2013) for programmatic research identifying the experiential drivers of team construct emergence. In other words, future research should identify the types of fine-grained team member interactions that give rise to emergent team properties.

Moreover, gaining a clear grasp of the events that enable team construct emergence requires research settings, such as laboratory experiments, where researchers have access to tools (e.g., audio and visual recordings) that provide a high level of control and a constant stream of environmental and behavioral markers. A fruitful line of inquiry in the laboratory would combine these behavioral and environmental traces with self-report measures of emergent team properties and our analytic approach in order to isolate the factors that catalyze emergent team properties. Marks et al. (2001) proposed a loop between processes and emergent states, such that one influences the other. Possibly, processes influence both the *levels* of less observable emergent phenomena as well as their actual emergence as *recognizable* properties. For example, information sharing may enable the emergence of trust as members enact behaviors that allow demonstrations (or violations) of trustworthy behavior.

In conclusion, developing a better understanding of the structure, function, and measurement of emergent team properties is crucial to the organizational sciences as the nature of work becomes more interdependent (Wood & Hoffman, 2010) and world-changing innovations rely on collaborative, interdisciplinary efforts (Uzzi, Mukherjee, Stringer, & Jones, 2013). Moreover, given their importance and complexity, specifying theoretic conceptions of emergent team properties *prior to measuring* these constructs is an essential step in rigorous research on teams (Kozlowski & Klein, 2000). The results of our two studies demonstrate that the measurement of emergent team properties is a complex affair, involving dynamic processes that can change the properties of measurement instruments over time.

Notes

1. All ME/I analyses here were also conducted using the single-level IRT model, excluding the isomorphic team calibration effects. LR tests and effect sizes based on Meade (2010) were consistent with those presented here. Details on these analyses are available by contacting the first author.

2. The MHRM algorithm occurs in three stages. The first stage is used to refine default initial values of parameter estimates; we used 2,000 cycles to refine these initial estimates. In the second stage, an expectation-minimization algorithm is used to further refine these estimates; we used 1,000 iterations to further refine initial estimates. Finally, in the third stage, the Markov Chain Monte Carlo–based MHRM procedure is conducted; we set the number of cycles for MHRM to 20,000 to ensure convergence. All models estimated here converged. For the MHRM Monte Carlo simulations, the sample size was set to 25,000. These values are the default values used in the flexMIRT program. Burnin and thinning were set to 10, which are the default values for flexMIRT. Performance of the MHRM was evaluated by monitoring acceptance rates, which stayed very close to .50 in all analyses (see Houts & Cai, 2013, for more information on the MHRM estimation algorithm).

3. The ICC for observability ratings of Study 1 variables was .56 for absolute agreement and .73 for consistency.

4. Model comparison results may be obtained by contacting the first author.

References

Adams, B. D., Thomson, M. H., Brown, A., Sartori, J. A., Taylor, T., & Waldherr, S. 2008. *Organizational trust in the Canadian forces*. Report no. CR-2008-038, Defence Research and Development Canada, Toronto.

Allport, F. H. 1967. A theory of enestruence (event-structure theory): Report of progress. *American Psychologist*, 22: 1-24.

Ambady, N., Bernieri, F. J., & Richeson, J. A. 2010. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, vol. 32: 201-271. New York: Academic Press.

Arthur, W., Jr., Bell, S. T., & Edwards, B. D. 2007. An empirical comparison of the criterion-related validities of additive and referent-shift operationalizations of team efficacy. *Organizational Research Methods*, 10: 35-58.

Baumeister, R. F., & Tice, D. M. 1988. Metatraits. *Journal of Personality*, 56: 571-598.

Bunderson, J. S., & Sutcliffe, K. M. 2002. Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects. *Academy of Management Journal*, 45: 875-893.

Cai, L. 2010. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35: 307-355.

Cai, L. 2013. flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., & Hansen, M. 2013. Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66: 245-276.

Cai, L., Thissen, D., & du Toit, S. H. C. 2011. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: SSI.

Campbell, D. T. 1955. The informant in quantitative research. *American Journal of Sociology*, 60: 339-342.

Chen, G., Mathieu, J. E., & Bliese, P. D. 2004. A framework for conducting multilevel construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes*, vol. 3: 273-303. Oxford, England: Elsevier.

Chiocchio, F., & Essiembre, H. 2009. Cohesion and performance: A meta-analytic review of disparities between project teams, production teams, and service teams. *Small Group Research*, 40: 382-420.

Cronin, M. A., Weingart, L. R., & Todorova, G. 2011. Dynamics in groups: Are we there yet? *The Academy of Management Annals*, 5: 571-612.

D'Innocenzo, L., Mathieu, J. E., & Kukenberger, M. R. in press. A meta-analysis of different forms of shared leadership–team performance relations. *Journal of Management*. doi:10.1177/0149206314525205

Drasgow, F., Levine, M. V., & Williams, E. A. 1985. Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38: 67-86.

Drasgow, F., Levine, M. V., & Zickar, M. J. 1996. Optimal detection of mismeasured individuals. *Applied Measurement in Education*, 9: 47-64.

Edmondson, A. C. 1999. Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44: 350-383.

Ellis, A. P. 2006. System breakdown: The role of mental models and transactive memory in the relationship between acute stress and team performance. *Academy of Management Journal*, 49: 576-589.

Fox, J. 2008. *Applied regression analysis and generalized linear models* (2nd ed). Thousand Oaks, CA: Sage.

Ghiselli, E. E. 1964. *Theory of psychological measurement*. New York: McGraw-Hill.

Goldstein, J. 1999. Emergence as a construct: History and issues. *Emergence*, 1: 49-72.

Hackman, R. J. 2012. From causes to conditions in group research. *Journal of Organizational Behavior*, 33: 428-444.

Hollingshead, A. B. 1998. Communication, learning, and retrieval in transactive memory system. *Journal of Experimental Social Psychology*, 34: 423-442.

Houts, C. R., & Cai, L. 2013. *flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Humphrey, S. E., & Aime, F., 2014. Team microdynamics: Towards an organizing approach to teamwork. *Academy of Management*, 57: 327-352.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. 2005. Teams in organizations: From I-P-O models to IMOI models. *Annual Review of Psychology*, 56: 517-543.

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

Jones, L. V., & Thissen, D. 2007. A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, 26: Psychometrics*: 1-27. Amsterdam: North Holland.

Kanawattanachai, P., & Yoo, Y. 2007. The impact of knowledge coordination on virtual team performance over time. *Management Information Systems Quarterly*, 31: 783-808.

Kankaras, M., Vermunt, J. K., & Moors, G. B. D. 2011. Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods and Research*, 40: 279-310.

Kozlowski, S. W. J., & Chao, G. T. 2012. The dynamics of emergence: Cognition and cohesion in work teams. *Managerial and Decision Economics*, 33: 335-354.

Kozlowski, S. W., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. 2013. Advancing multilevel research design capturing the dynamics of emergence. *Organizational Research Methods*, 16: 581-615.

Kozlowski, S. W. J., & Klein, K. J. 2000. *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco: Jossey-Bass.

LeBreton, J. M., & Senter, J. L. 2008. Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11: 815-852.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. 2008. A meta-analysis of teamwork processes: Tests of a multidiemsional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61: 273-307.

Lewis, K. 2003. Measuring transactive memory systems in the field: Scale development and validation. *Journal of Applied Psychology*, 88: 587-604.

MacCorquodale, K., & Meehl, P. E. 1948. On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55: 95-107.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. 2001. A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26: 356-376.

Mathieu, J. E., & Marks, M. A. 2006. *Teamwork process items*. Unpublished scale, University of Connecticut.

Mathieu, J., Maynard, 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: 410-476.

McAllister, D. J. 1995. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38: 24-59.

Meade, A. W. 2010. A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, 95: 728-743.

Meade, A. W., & Wright, N. A. 2012. Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, 97: 1016-1031.

Mesmer-Magnus, J. R., & DeChurch, L. A. 2009. Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology*, 94: 535-546.

Mohammed, S., Tesler, R., & Hamilton, K. 2012. Time and shared cognition: Towards greater integration or temporal dynamics. In E. Salas, S. Fiore, & M. Letsky (Eds.), *Theories of team cognition: Cross-disciplinary perspectives*: 87-116. New York: Taylor and Francis Group.

Moreland, R. L., & Myaskovsky, L. 2000. Exploring the performance benefits of group training: Transactive memory or improved communication. *Organizational Behavior and Human Decision Processes*, 82: 117-133.

Morgeson, F. P., & Hofmann, D. A. 1999. The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Journal*, 24: 249-265.

Reise, S. P., & Waller, N. G. 1993. Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, 65: 143-151.

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34: 100-114.

Stark, S., Chernyshenko, O. S., & Drasgow, F. 2006. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91: 1292-1306.

Tesluk, P., & Mathieu, J. E. 1999. Overcoming roadblocks to effectiveness: Incorporating management of performance barriers into models of work group effectiveness. *Journal of Applied Psychology*, 84: 200-217.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. 2013. Atypical combinations and scientific impact. *Science*, 25: 468-472.

Wood, L., & Hoffman, B. J. 2010. *The changing nature of work: A meta-analysis*. Paper presented at the 25th annual meeting of the Society for Industrial and Organizational Psychology, Atlanta.

Yang, J. S., Monroe, S., & Cai, L. 2012. *Multiple group multilevel item bifactor analysis*. Paper presented at the 74th annual meeting of the National Council on Measurement in Education, Vancouver.

Zickar, M. J. 1998. Modeling item-level data with item response theory. *Current Directions in Psychological Science*, 7: 104-109.