Team Data Analysis Using FATE: Framework for Automated Team Evaluation

Alec Ostrander, Stephen Gilbert, & Michael Dorneich

VRAC, Iowa State University, 1620 Howe Hall, Ames, IA 50011, USA
alecglen@iastate.edu, gilbert@iastate.edu, dorneich@iastate.edu

Abstract. In this paper we introduce a conceptual framework for the design of automated team evaluation processes (FATE), inspired by lessons learned from multiple intelligent team tutoring experiences. The framework consists of five phases. The first, Team Construct, defines the theoretical basis of the evaluation and therefore the end goal of the evaluation process. The second, Behavioral Markers, defines a method for identifying the otherwise unobservable constructs. The third, Raw Data, defines the data to be captured and recorded. The fourth, Enriched State Representation, defines a method for making the data directly relevant for team evaluation. The fifth, Team Metric, is the end goal of the evaluation defined by team constructs and derived from the enriched state representation. The framework is organized in a "V" shape to act both as a hierarchical model relating teaming theory to scenario-specific data and as a sequential process flow diagram representing the steps recommended to design an ideal team evaluation process. Each phase of the framework is described in detail, and its use is demonstrated with a hypothetical emergency response training scenario.

Keywords: team assessment, intelligent team tutoring system, data analysis.

1 Introduction

A common challenge in team tutoring research is the difficulty of automated team evaluation [1]. Effective team tutoring requires meaningful performance metrics rooted in teaming theory, but the complexity of team dynamics often makes ideal assessment methods impossible or impractical. Depending on characteristics and pace of the task, evaluation may be difficult even for human coaches, let alone an intelligent team tutoring system. For example, an ideal metric for team communication might involve the timing and semantic content of verbal interactions between team members, but tools for reliable speech analysis are not yet widely available [2]. Instead, more creative data processing methods must be used to generate metrics that accurately reflect the intended attributes of team performance.

Eliciting team metrics from task performance data is not a trivial task, especially if the data are incomplete or ambiguous. The relevant task data must first be identified and transformed to synthesize a more meaningful representation of the complex team task space. Furthermore, team metrics generated from that representation must be vali-

dated to ensure they are measuring the intended team constructs as understood by human evaluators. Overall, team data analysis can be especially difficult to perform while keeping the larger picture in mind, leading to results that only describe one component of a team's performance. To alleviate this difficulty, there is need for a consistent process for automating team evaluation.

This research presents the Framework for Automated Team Evaluation (FATE) to guide the process of team evaluation. The conceptual framework describes the progression of a tutoring system from the initial design of a tutoring scenario through learner evaluation and is divided into five stages based on the type of information and the level of abstraction they represent. The following section reviews literature from team training and evaluation research that provides a theoretical basis for the framework. Section Three describes the framework in detail, highlighting the role each component plays and how they relate to each other to form a sequential process for evaluation design. Section Four demonstrates how the framework can be used to improve team evaluation outcomes, making use of a hypothetical team tutoring scenario inspired by the experiences of the authors. Lastly, Section Five concludes with a discussion of the framework's benefits, limitations, and applications.

2 Related Work

Intuitively, research in intelligent team tutoring borrows from work in team training and evaluation by humans. While intelligent team tutoring is still a nascent technology, there is a breadth of research digging into how human teams function. Salas, Stagl, Burke, and Goodwin [3] identified over 130 different frameworks in the literature that describe the way teams behave and perform in varying levels of specificity and complexity. The most general and overarching approach described is the input-process-output (IPO) model of teams, which characterizes a team as a functional system (the process) that when presented with a given context and the scenario (the input) will behave to produce a certain outcome (the output) [4]. Despite limitations, modeling teams in this fashion is parsimonious in the way it affords describing each factor separately and flexibly. However, while the IPO model is useful for understanding team behavior generally, it stops short of defining specific factors that make teams effective. It is left for other models to fill in that gap.

Several models of team performance define a core set of factors that make a team perform better or worse. In the present work, these factors will be termed team constructs. An influential review compiled by Salas, Shuffler, Thayer, Bedwell, and Lazzara [5] advocates for the existence of nine critical considerations, sometimes referred to as the "9 C's of teamwork." These include six core processes of a team – e.g., cooperation, coordination, and communication – and three external influencing conditions – e.g. task context and team composition. More recently, Sottilare, Burke, Salas, Sinatra, Johnston, and Gilbert [6] conducted a meta-analysis of the literature on teamwork, team performance, and intelligent team tutoring system applications with the intent to find evidence of causal relationships between different team behaviors and team performance and learning. The authors also documented strong effects of communication,

coordination, conflict management, leadership, and team cognition across multiple studies.

Although the team constructs mentioned above are useful for describing team performance at a general level, the means of observation are left ambiguous. How does one measure team coordination or team conflict? Behavioral markers offer an answer to this question. A behavioral marker is a real-world, objective behavior by an individual or team for which a relationship with an otherwise unobservable construct can be assumed [7]. Once this relationship is established, the frequency, quality, or other features of the marker can be used as a stand-in for measuring the construct one is interested in. Behavioral markers see widespread use in psychology and the social sciences and have been adopted for use in team evaluation. In particular, Sottilare, Burke, Salas, Sinatra, Johnston, and Gilbert [6] used a systematic process to develop behavioral markers for each of five team constructs. For example, behavioral markers for communication included "occurrences of task relevant information being shared" and "occurrences of team members providing verbal feedback to one another."

Using markers, a human tutor could evaluate a team by tallying and grading occurrences. However, before automation can perform the same evaluation, the markers must be formally and quantitatively specified. To this end, previous works have developed their own context-specific methods to measure team performance automatically [8, 9, 10, 11]. Some simply substitute overall task performance, while others have intentionally focused on the measurement of team constructs. For example, Cooke, Salas, Cannon-Bowers, and Stout [12] reviewed several methods to identify team cognition, while Stevens, Galloway, Berka, and Sprang [13] used an electroencephalogram (EEG) to measure cognitive synchrony between team members as a marker of team cognition and cohesion. Bowers, Jentsch, Salas, and Braun [14] presented work on speech pattern analysis affecting team task performance that is relevant to communication. The proposed framework provides a method of characterizing each of these efforts and lays a path for the development of future evaluation methods.

3 Framework for Automated Team Evaluation

The previous sections laid out the building blocks for a conceptual framework to guide the development of team evaluation processes. An evaluation process should be designed with team constructs as the theoretical basis so that final results are meaningful in terms of the purpose of the evaluation. From this perspective, behavioral markers act as an intermediary step relating team constructs to the raw data collected, and an enriched state representation is used as an intermediary step to infer the occurrence of behavioral markers from the data. The complete framework is modeled as five distinct phases across three levels of abstraction, with the intent being that they be implemented in sequential order. This model is illustrated in Fig. 1 below, and the following subsections detail the purpose and usage of each phase.

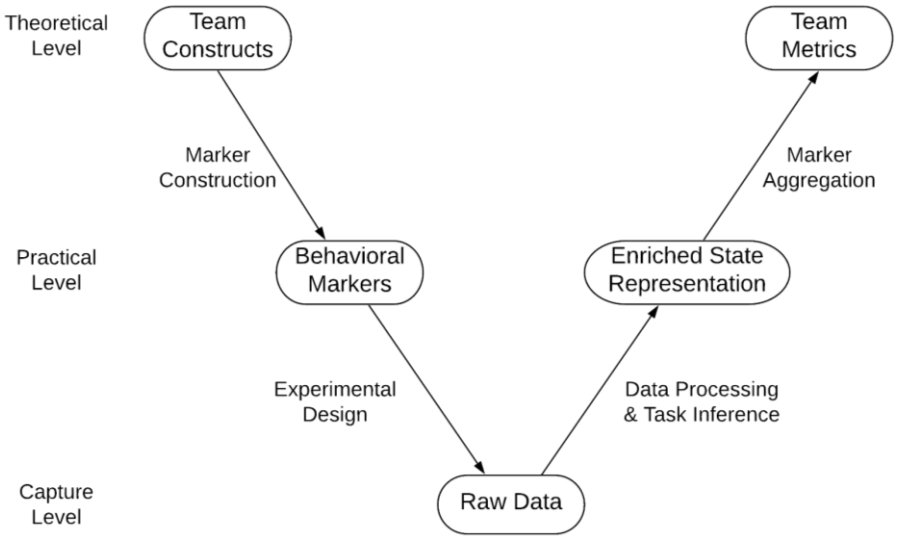

Fig. 1. Illustration of the Framework for Automated Team Evaluation (FATE), demonstrating its duality as both a hierarchical model and a sequential process. At the theoretical level, an intelligent team tutoring system should be designed with explicit team constructs in mind. Since evaluating the tutor's effect on those constructs is a priority, the end goal is one or more metrics that directly represent the team's performance on those constructs. Getting from start to finish requires traveling down the "V" to define the data that should be collected, and back up the "V" to design the processes needed to generate team metrics.

3.1 Team Constructs

As stated previously, *Team Constructs* are the abstract behavioral or social concepts that one designing a team tutoring system wishes to improve in the teams using the system. They are key to the motivation of designing the tutor, but there is generally no simple, direct measure that would make evaluation trivial. Examples are team trust, shared cognition, coordination, and communication. There is a large breadth of literature describing constructs that may be strived for to achieve more effective team performance. A researcher investigating team performance differences between co-located and distributed teams, for example, might choose to focus specifically on the construct of team communication. While choosing the construct of interest is useful to narrow the focus of the work and design a more targeted evaluation, the actual method of assessing the communication ability of the team remains unclear.

3.2 Behavioral Markers

In contrast to theoretically important but practically unusable team constructs, *Behavioral Markers* are well-defined, observable behaviors. The purpose of a behavioral marker is to, either individually or in combination with other markers, serve as a proxy for the team construct one is interested in evaluating. When using behavioral markers

for a team evaluation, it is critical that sufficient attention is given to verifying the relationship between the markers chosen and the construct they are intended to represent. For example, to evaluate team communication ability for the previous example, behavioral markers might be:

1. Occurrences of team members updating each other on goal progress
2. Occurrences of verbal acknowledgement and/or feedback

These are measurable events and therefore bring the researcher closer to being able to perform a meaningful team evaluation. In fact, this may be all that is needed for a human tutor. However, enabling automated evaluation by an intelligent team tutoring system requires an additional step.

3.3 Raw Data

At the lowest level of abstraction for an automated evaluation process is the *Raw Data* phase. Raw data are the data that are actually captured from the team members and environment during an evaluation. These may include audio, video, screen capture, keystrokes, physiological measurements, questionnaire responses, etc. While the distinction between team constructs and behavioral markers is the theoretical possibility of direct measurement, the distinction between behavioral markers and raw data is characterized by practical limitations on measurement due to the technology available for evaluation. In the team coordination example, it would be reasonable for a human trainer to take note of the behavioral markers described in the previous paragraph, but the same task is well out of reach for all but the most advanced speech recognition and natural language processing systems. Instead, the behavioral markers should be used during experiment design and implementation to motivate what data is captured by the system. Because the behavioral markers of interest are both based on verbal interactions, it follows that audio should be the primary data captured by the system.

While the division between behavioral markers and raw data could also be made for individual evaluation, it is especially relevant for teams. This is because, when designing the evaluation for an individual tutoring system, the designer can leverage the fact that the system need only assess the learner's interaction with the system itself and limit the modes of interaction to those that are easily interpreted computationally. The team tutoring system designer is not so fortunate, as the complexities of human-human interaction that are inherent to teams must be included in a complete assessment.

3.4 Enriched State Representation

The reason for defining the step from behavioral markers to raw data is that it provides a bridge from behaviors that can technically be observed to what the system is capable of observing. Once data has been captured, however, the raw data must be processed into a form from which behavioral markers can be inferred. This is referred to as an *Enriched State Representation*. Whereas raw data may represent the state of the team from a computational or individual perspective, this phase is the state of the team from

a task perspective. The term “enriched” is used because the process of generating the new representation often requires using knowledge of the task structure and inference to supplement the otherwise incomplete data.

The value of an enriched state representation is that it enables highly flexible, meaningful analysis of the data in terms of the behavioral markers it was meant to capture. Just as it was vital to ensure the derived behavioral markers accurately represented the intended construct, validation of the process to move from raw data to an enriched representation is essential to ensure it yields an accurate record of the behavioral markers. Therefore, the more complex the scenario, and the more types of data collected, then the more important the processing step from data to enriched representation.

Consider again the scenario of evaluating team communication through the behavioral markers of team progress updates and acknowledgement/feedback, given only raw audio data. The data itself cannot indicate either behavioral marker – a change in audio volume or frequency does not mean a progress update. However, more complex inference, combined with knowledge of the training task and environment, might be able to make sense of patterns in the data and provide stronger insights. For example, an enriched representation could combine a several-second increase in volume from Team Member 1 with knowledge that Team Member 1 had just completed a goal task and thus infer that Team Member 1 was communicating a progress update. If a shorter-length volume increase occurred from Team Member 2 immediately following the progress update, the enriched representation has good reason to infer an acknowledgment at that time. The key characteristics of an enriched representation are that it portrays the team’s behaviors in terms of the task structure and that it can be used to identify occurrences of behavioral markers.

3.5 Team Metrics

The fifth and final phase in the team evaluation process is the *Team Metric*. If the goal of the evaluation is to understand the team’s behavior in terms of one or more team constructs, then the team metric is the evaluation model’s optimal quantitative representation of those constructs. Conveniently, team metrics can be derived intuitively from the enriched state representation by identifying and aggregating the corresponding inferred behavioral markers. For example, once the behavioral markers of progress updates and acknowledgment/feedback have been represented in the enriched representation, summing these markers over the course of the tutoring session could provide an easy, quantitative measurement of the team’s communication ability.

4 Demonstration of the Framework

This section demonstrates how to use the framework to guide the design of a team evaluation process. The subject of evaluation is a hypothetical emergency response training. In this scenario, two learners participate in a desktop virtual simulation by commanding first-person avatars in the scenario. The simulation begins with the avatars as first responders approaching the scene of a two-vehicle accident on a busy highway.

Team members should communicate to come to a shared understanding of the situation, form a plan, and perform actions leading to optimal scenario outcomes. The scenario is designed such that one victim in each vehicle is injured, and that one vehicle will begin to catch fire if not extinguished early. Time pressure is applied, and effective teamwork is paramount to the scenario's successful resolution.

Following the framework, the evaluation design process begins by selecting team constructs since they are what the system is trying to train. In the case of emergency response teams, the designer may decide to focus on coordination as the most relevant factor to task performance. This is reflected in the scenario design, and along with a starting point, it defines the end goal of the evaluation process. However, as described in Section 3, it is still unclear at this point how coordination can be assessed.

To decide how coordination is best assessed, the framework advises the construction of behavioral markers – specific team or task events that can be linked as observable indicators of the chosen constructs. While task-generic team markers have been identified in [6], greater specificity may be achieved by considering what successful performance of the construct means in terms of the specific task at hand. To begin, explicitly list the high-level actions team members may perform. For the emergency response scenario, these might include:

- Assess the environment
- Block off lanes of traffic
- Assess conditions of victims
- Extinguish possible fires
- Evacuate victim from vehicle
- Perform first aid
- Request medical assistance

With actions defined, any instance in which both team members are performing one of the actions could technically be considered a marker of coordination. However, finer specification can lead to greater evaluation accuracy. Some examples of context-specific markers could include:

- Occurrence of team members splitting up to triage each vehicle simultaneously
- Occurrence of team members simultaneously evacuating victims and extinguishing fire once the vehicle ignites.
- (Marker of poor coordination) Occurrence of unnecessary duplicate actions on the same object/entity by both team members.

At this point, a human tutor could take the set of behavioral markers as defined and note their occurrence or lack thereof over the course of the simulation. However, automating an intelligent tutoring system to perform the same assessment requires a more formal specification.

To guide that transition, the next stage in the framework finally reaches the level of the actual system with the raw data phase. The intent is that the designer uses the previously constructed behavioral markers as rationale for what data the system should

capture. In the emergency response training, the main data captured should be simulation states and events since the behavioral markers all relate to actions taken. For example, the first marker above motivates that each avatar's position in the environment should be recorded. The second marker motivates that the action currently being performed should be recorded, if any. The third marker adds the need to record not just the action but also any objects on which it is being performed.

It is interesting to note that there is no need for audio recordings in this evaluation design. Although audio to capture team interactions is a staple of team evaluation, it is not motivated by the goals of the current evaluation. That is not to say it cannot or should not be recorded; an audio recording may be valuable to gain other insights such as verifying individual intent later on. However, it does not have a role in the formal evaluation as designed and trying to add it without clear reason can easily lead to confusion and unneeded complexity.

Once the data to be captured by the system has been specified, the next step is to construct an enriched state representation. The goal in this step is to transform the data from a system-level, individual-centric perspective to a team-focused perspective from which occurrences of behavioral markers can be found. In this example, the transformation is from a time series of simple positions and events to a timeline of which events are overlapping when, termed *interactions*. The enriched representation could be used to say, "At time 32.5s, the learners began coordinating by splitting up and triaging separate vehicles." With this step of the process implemented, identifying behavioral markers in the timeline is as trivial as searching occurrences of the relevant interaction.

While in this case the transformation into an enriched representation was fairly simple, without careful design and forethought it can easily become very difficult or impossible. If the required raw data are not all captured, a simple transformation is no longer sufficient, and inferential or statistical methods may be required to reach a useful representation. In addition, in cases where audio data is deemed a necessary part of data capture, the processing can be highly complex and tedious.

However, once an enriched state representation has been developed, the process of team metric derivation is trivial. By searching through the new representation, counts of each behavioral marker can be accumulated. These counts, or a derivative such as frequency, subsequently act as the metric for the team construct to which the behavioral marker corresponds. If multiple markers are used to evaluate one construct, it is recommended that Cronbach's alpha or other reliability methods are used to validate the internal consistency of the final metric.

5 Conclusions

This paper presented and demonstrated the use of the Framework for Automated Team Evaluation (FATE) for the development of team evaluation processes for intelligent team tutoring systems. The framework encapsulates and organizes knowledge and lessons learned by the authors as a result of several team tutoring experiences. As a general

design guide, the proposed method does not provide specific instructions or recommendations for process design. However, it is formulated to be applicable to a wide range of different potential team tutoring scenarios and systems.

6 References

1. Bonner, D., Gilbert, S., Dorneich, M. C., Winer, E., Sinatra, A. M., Slavina, A., Macallister, A., Holub, J. The challenges of building intelligent tutoring systems for teams. In Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting, pp. 1981–1985. Washington, DC (2016).
2. Sinatra, A. M., Gilbert, S., Dorneich, M., Winer, E., Ostrander, A., Ouverson, K., Johnston, J., Sottilare, R. Considerations for dealing with real-time communications in an intelligent team tutoring system experiment. In Proceedings of the Assessment and Intervention during Team Tutoring Workshop,pp. 28–35 (2018)
3. Salas, E., Stagl, K. C., Burke, C. S., Goodwin, G. F. . Fostering team effectiveness in organizations: toward an integrative theoretical framework. Nebraska Symposium on Motivation. Nebraska Symposium on Motivation, **52**, pp. 185–243 (2007).
4. Mohammed, S., Hamilton, K.. Input–process–output model of team effectiveness. In S. G. Rogelberg (Ed.), Encyclopedia of Industrial and Organizational Psychology, Vol. 1 pp. 354–355. Thousand Oaks, CA: SAGE Publications, Inc. (2007).
5. Salas, E., Shuffler, M., Thayer, A., Bedwell, W., Lazzara, E. Understanding and improving teamwork in organizations: A scientifically based practical guide. Human Resource Management, **54**(4), 599–622 (2015).
6. Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J. H., Gilbert, S. (2017). Designing Adaptive Instruction for Teams: A Meta-Analysis. International Journal of Artificial Intelligence in Education, **45** (2017).
7. Flin, R., Martin, L. Behavioral markers for crew resource Management: A review of current practice. International Journal of Aviation Psychology, **11**(1), pp. 95–118 (2001).
8. OECD. (2017). PISA 2015 Assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving (revised ed). Paris: OECD Publishing.
9. Ostrander, A., Bonner, D., Walton, J., Slavina, A., Ouverson, K., Kohl, A., Gilbert, S., Dorneich, M., Sinatra, A., Winer, E. Evaluation of an intelligent team tutoring system for a collaborative two-person problem: Surveillance. Computers in Human Behavior. (2019).
10. Walton, J., Gilbert, S., Winer, E., Dorneich, M., Bonner, D. Evaluating distributed teams with the team multiple errands test. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), pp. 1–12 (2015).
11. Zachary, W., Cannon-bowers, J. A., Bilazarian, P., Krecker, D. K., Lardieri, P. J., Burns, J. The advanced embedded training system (AETS): An intelligent embedded tutoring system for tactical team training. International Journal of Artificial Intelligence in Education, **10**, pp. 257–277 (1998).
12. Cooke, N. J., Salas, E., Cannon-Bowers, J. A., Stout, R. J. Measuring team knowledge. Human Factors: The Journal of the Human Factors and Ergonomics Society, **42**(1), 151–173 (2007).
13. Stevens, R. H., Galloway, T., Berka, C., Sprang, M. Can neurophysiologic synchronies provide a platform for adapting team performance? Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5638 LNAI, pp. 658–667 (2009).

14. Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing Communication Sequences for Team Training Needs Assessment. Human Factors: The Journal of the Human Factors and Ergonomics Society, 40(4), 672–679.