Objective Metrics for Evaluation of Collaborating Teams

David Noble and John Kirzl [*]
Evidence Based Research, Inc.
1595 Spring Hill Road, Suite 250
Vienna, VA 22182-2216
703-287-0312
noble@ebrinc.com

Abstract

Collaboration evaluation has two principal goals. First, it seeks to quantify changes in team performance, in order to determine the extent to which a new technology, process, or organization improves team effectiveness. Second, it seeks to explain the reasons for changes in effectiveness.

This paper outlines an approach for achieving both of these goals using collection procedures which reflect the evaluation sponsor's cost requirements and collection constraints. The paper begins by discussing the advantages of objective performance measures. It then describes measures for three important collaboration products: situation understandings, plans, and decisions. It concludes by describing how to create a quantitative audit trail for documenting the reasons for a new technology, process, or organization's impact.

1. Objective Performance Measures

Evaluators use objective measures to quantify performance. For example, they quantify situation understanding by asking participants questions about the situation. They then "grade" the answers to the questions using an answer key generated by subject matter experts.

Objective performance measures contrast with subjective measures that are based on the opinions and self assessments of the evaluation participants. For example, a common subjective measure of situation understanding is the extent to which a participant says he or she understands the situation. Another is the extent to which participants say they like or dislike the information technology supporting the situation understanding.

Surprisingly, many evaluations of new technologies, processes, or organizations use subjective measures exclusively. While subjective measures are an important element of an assessment (it is very important to know when people don't like a product or feel it doesn't contribute to a better understanding), they should not be the sole basis for an evaluation because, as documented in the JFCOM Presentation LOE and elsewhere, people's self assessments do not always align with performance.

For this reason, it is important to include objective performance measures when evaluating the impact of new tools, processes, or organizations on collaboration and team performance.

[*] Work sponsored by DARPA and the Office of Naval Research

Objective measures provide the best evidence for sponsors or skeptics. In addition, they help document a credible casual audit trail able to explain the reasons for the performance impact.

Though data collection to support objective measures is often more labor intensive than data collection for subjective measures, it does not need to be costly and intrusive. In fact, using some of the methods outlined in this paper, collecting data for objective measures can be less intrusive than collecting data to support subjective opinions.

2. Methodology Background

The methodology described here builds on twenty years of EBR experience employing objective measures. Some of these previous efforts include:

- HEAT (DCA 1980-1981)
- ACCES (ARI 1985)
- ACCES Reinvented (1989)
- NATO COBP for C2 Assessment (1991-1998)
- CPOF C2 Experiments (1999-2003)
- JFCOM C2 Experiments (2000-2003)
- Revised NATO COBP (2001-2002)
- Collaboration Metrics (ONR 1999-2003)
- NCW Conceptual Framework Metrics (2002-)

The first of these, the HEAT (Headquarters Effectiveness Assessment Tool) pioneered the use of objective measures. The HEAT methodology builds on a model of information processing and flow (Figure 1) within a headquarters. This flow generalized the OODA (Observe-Orient-Decide-Act) loop originally based on pilot cognitive tasks to a more general process focused on headquarters activities. The HEAT methodology developed metrics for each of the six steps in Figure 1.

The current methodology traces its parentage to this foundational work. However, the newer methodology benefits from years of practical evaluation experience and from significant theoretical advances in understanding the cognitive mechanisms underlying situation assessment, decision making, collaboration, and recent Command and Control concepts, such as Network Centric Warfare.

3. Objective Measures of Product Quality

When the objective of the team is to create an intellectual product (e.g., a situation assessment, plan, or decision), then the quality of this product is the key measure of team effectiveness. A team that produces a good quality product on time is performing well, no matter how chaotic its internal processes may seem to outsiders. Conversely, a team that appears to function smoothly but produces a poor quality product is not performing well.

The objective quality measures "score" a product by grading its contents using an answer key developed by subject matter experts. Thus, the measures resemble a teacher's grade on an essay

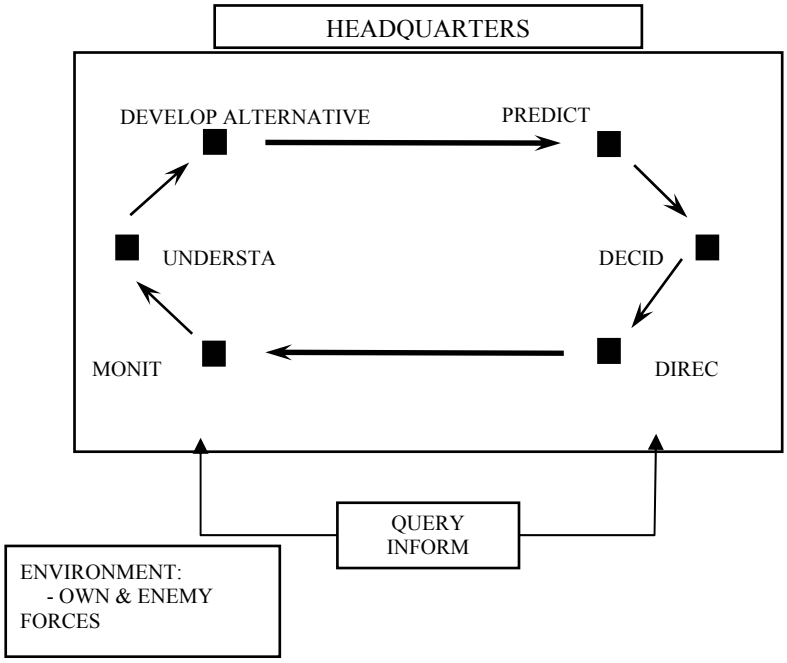

Figure 1. The HEAT Analytic Structure

test. In assigning the grade, the teacher has (at least in his or her mind) a check list of points that should be addressed, additional checklists of correct and incorrect facts that students often mention. The test is then graded by assigning points for correct items mentioned, not adding points for omissions, and subtracting points for incorrect statements.

Though answer keys for each kind of product always address the same kinds of issues appropriate to that kind of product, a specific answer key must be developed separately for the context that the product addresses.

Three of the most important intellectual products are situation assessments, plans, and decisions.

3.1 Objective Measures for Situation Assessment

Situation assessments are understandings of the external situation. In military teams, a situation assessment addresses all aspects of the external situation relevant to mission success. Situation assessment includes the identity and location of hostile and friendly units and of non-participants. However, a situation assessment goes beyond"who's where" to probe situation understanding. This additional understanding concerns knowledge of the status and capabilities of forces, identification of opportunities and risks, and projections of possible future actions.

Some examples of questions that are asked to evaluate a situation assessment include:

1. Who is doing y? ("y" is an activity")
2. What are the actions he is most likely to do next?
3. What are the situation factors to be most concerned about now?
4. What are the situation factors that might have the greatest impact on y?

5. How certain is each of these factors?
6. What information is able to resolve these uncertainties?
7. What are our greatest risks and opportunities?
8. What are the different possible ways that x and can react to y?
9. What will be the first cues that that x is responding in each of these ways?

The most straight-forward method for collecting data for situation assessment is to ask evaluation participants about each of these issues. A less intrusive method of obtaining the needed assessment data is for the participant playing the role of commander to ask his team members for their assessments or to ask them to prepare a briefing on assessments. Finally, if this is not possible, the evaluators may estimate the level of situation understanding non-intrusively by recording overheard conversations, reviewing prepared documents, and by observing behavioral indicators of situation understanding.

In fact, non-intrusive data collection can work very well, and has been employed extensively for many different kinds of evaluations. Figure 2 summarizes situation scores for numerous understanding for numerous exercises over the past fifteen years. All of these data were collected non-intrusively, either from overheard statements during planning or by assessing the content of situation briefings

Surrogate Baseline	**Day 1**	**2**	**3**	**4**	**5**	**Overall**	
Joint Command Exercise in the 1980's	.52	.62	.52	.6	.65	.6	Median % Correct
U.S. Army Divisions circa 1990 – 1993	.74	.82	.84	.82	.82	.81	Average % Correct

Surrogate Baseline	**.Week 1**	**Week 2**	**Week 3**	**Overall**	
Joint Command Exercise in the 1980's	.93	.86	.69	.83	Average % Correct

Spiral 3	**Day 1**	2	**3**	**4**	**5**	**6**	**7**	**Overall**	Weighted % Correct
	.87	.85	.90	.71	.58	.97	1.	.86	

MCO2	**Week 1**	**Week 2**	**Week 3**	**Overall**	Weighted % Correct
	.78	.85	.87	.85	

Figure 2. Examples of Situation Assessment Scores obtained noninstrusively.

3.2 Objective Measures for Plans

EBR has employed three types of objective measures for plan quality measure. The first avoids any requirement for subjective assessment on quality. Instead it measures the useful life of the plan compared to its intended life. No plan "survives contact with the enemy," but better plans last longer.

The second type of measure requires an assessment key that the plan evaluators can create by reviewing mission orders and commander discussions. For example, by reviewing orders and discussions, plan evaluators can enumerate all of the commander's objectives. They can then measure a quality of the plan as the fraction of commander's objectives that plan addresses.

A third type of measure requires an expert's answer key that enumerates the issues that a good plan should address. Because these issues may not have been made explicit in command orders, their enumeration requires expert judgment. An example is the number of important plan contingencies that the plan addresses.

In summary, three objective measures for plan quality are:

1. The useful life of the plan compared to its intended life
2. The fraction of commander's objectives addressed by the plan
3. The number of important contingencies that the plan addresses

3.3 Objective Measures for Decisions

There are two classes of metrics for decision quality: 1) the extent that the decision maker considers key factors, and 2) expert rating of the alternative selected. EBR does not use decision outcome as an objective measure for decision quality because a large number of factors in addition to the decision impacts outcome. In addition EBR does not evaluate decision quality from the decision process employed, because many of the best expert decisions are recognition-based, and do not use or need more deliberate processes of formally enumerating and evaluating multiple alternatives [Zsambok, 1993].

The first class of metric is the extent that the decision maker considers key factors: e.g., consideration of situation drivers such as centers of gravity, hedging for critical uncertainties, and keeping options open. This is an "anti-blindside" measurement, based on the finding that decisions which overlook some key aspect of the problem are at risk for leading to very bad out outcomes while those that do not overlook key aspects usually lead to good outcomes.

The second class of metric is expert rating of the alternative selected. Experts working in familiar situations know the different ways that a problem can be attacked, and know the strength and weaknesses of each of these different ways. They know, for example, that a particular course of action is highly risky because an adversary has an effective countermove. To employ this measure, at the completion of the participants' activities evaluators pick key decision points that occurred during mission execution. They note the situation information available to the decision maker at each of these the points, and taking that into account rate each of the different plausible options available at that time. The score of decision quality is the rating of the expert-rated option most similar to what the decision maker selected.

4. **Understanding Reasons for Product Quality**

Usually a sponsor desires to understand not only whether a new tool, process, or organization is improving team performance and product quality, but also to understand the reasons for the improvement. Equally important, the sponsor needs to know expected improvements were not observed.

Explanatory audit trails can identify the reasons for a changed environment's impact on team performance and product quality. EBR does this by documenting changes to two critical intervening variables: team members' understandings and team behaviors. Doing this requires objective metrics for measuring team member's understandings and behaviors. It also requires a model that describes the connection between an environment change, changes in the evaluation participant's understandings, changes in behavior, and product quality or action effectiveness.

4.1 When a Single Person Produces the Product

Figure 3 is a top level view of a model that EBR has employed for understanding the impact of new information technology on the quality of situation understanding and decision making by experienced decision makers working in familiar circumstances [Noble, 2000]. This model posits that experienced people identify promising actions to deal with a problem because they know the general sorts of action that work in various kinds of situations, they know the situation properties (usually abstract) of the situation that must be true for an action to work and be appropriate, and they know how to infer these abstract properties from concrete observable features. For example, in air defense, criteria for destroying an incoming track are that the track be hostile, intending to attack, and capable of causing harm. Experienced people know the different ways of destroying a target and know how to infer the abstract properties (e.g., intends to attack) from the concrete observable properties.

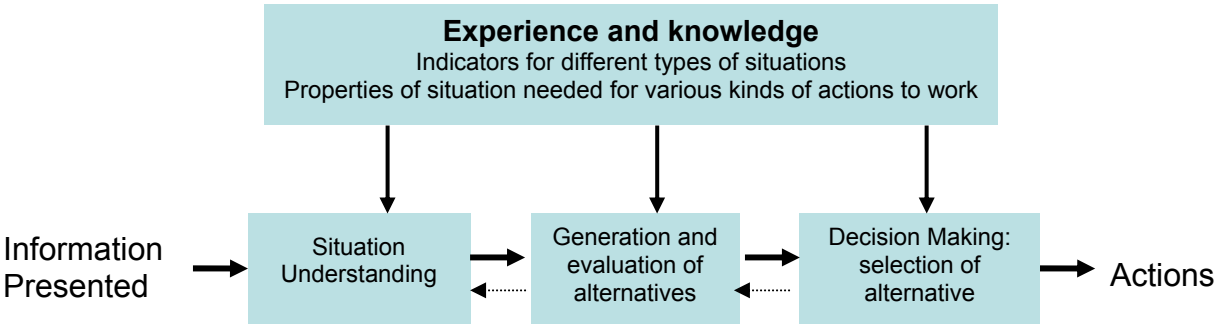

Figure 3. A Model for Situation Understanding, Alternative Generation, and Alternative Selection

In this model, a key part of situation understanding is inferring the abstract properties a situation must have in order for an action to work and be appropriate, the generation of alternatives is accomplished by recalling from memory and shaping the standard alternatives for this situation, and alternative evaluation is accomplished by determining the extent to which a particular situation has these desired properties.

EBR employed this model to understand the reasons for the situation understanding results obtained in one of the experiments in DARPA's Command Post of the Future. That evaluation compared the results among three different treatments (Figure 4): Baseline, with traditional color coding to represent force allegiance; Treatment B, with traditional color coding supplemented with "blobs" to help people infer aggregate force properties such as relative force strength at particular locations, and Treatment A, which tried out a new color coding scheme in which colors represented force structure function (e.g., artillery) and icon shape coded for allegiance. Thus, in Treatment A both friendly and adversary artillery had the same color. Opposing forces were differentiated by icon shape.

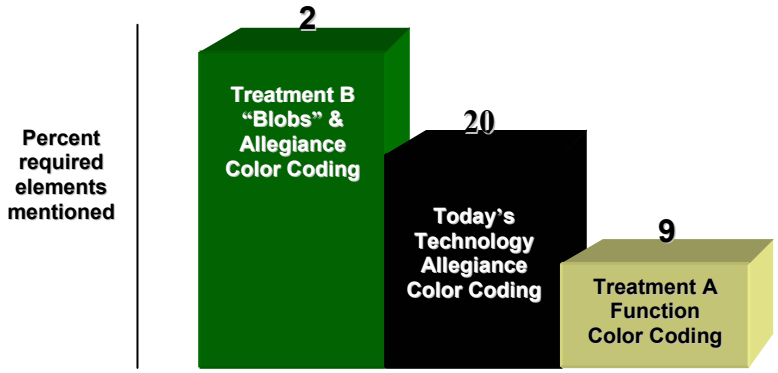

Figure 4. Evaluation of Situation Understanding in CPOF Limited Objective Experiment

In this experiment, the blobs in treatment B significantly improved situation understanding. The new functional color coding, however, reduced the quality of situation understanding. In this particular experiment, this reduction likely occurred because evaluating relative force strength was important to the overall assessment, and the non-conventional coding made this harder. Note that this result applied in this particular context. In another context, where the assessment required quickly assessing unit function, treatment A might be the more effective treatment.

Though this model is very simple, it is very helpful in understanding the experiment results. It also helps in knowing how to generalize the results

4.2 When a Team Produces the Product

When teams of people work together to create a product, then documenting an explanatory audit trail needs a somewhat more general model than when an individual works alone to create a product. EBR developed the model it uses as part of an Office of Naval Research SBIR in collaboration metrics [Noble, 2002]. Figure 5 summarizes the model framework, and Figure 6 and 7 convey different aspects of the model.

The aspect shown in Figure 6 emphasizes the knowledge and understandings that team members need in order to work together effectively. These are the "knowledge enablers" that make effective collaboration and teamwork possible.

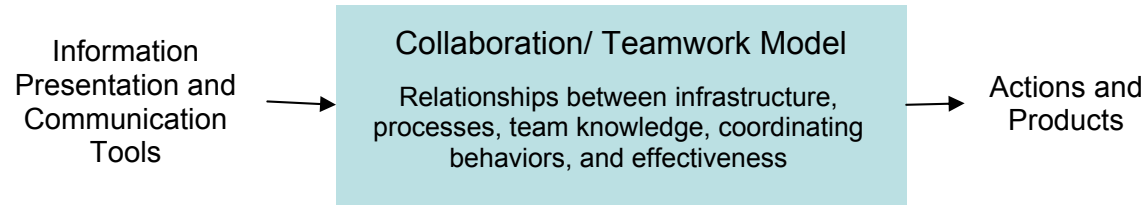

Figure 5. Framework for Collaboration/ Teamwork Model

This knowledge both enables and is enabled by the different kinds of team activities [Wegner, 1987]. Figure 6 includes three of these: team set up and adjustment, group problem solving, and synchronize and act. Team set up activities usually occur earlier and "synchronize and act" later,

but in most teams these activities re-occur as long as the team continues. Thus, most teams will revisit objectives, roles, and tasks as they solve problems and act together and discover need for clarification.

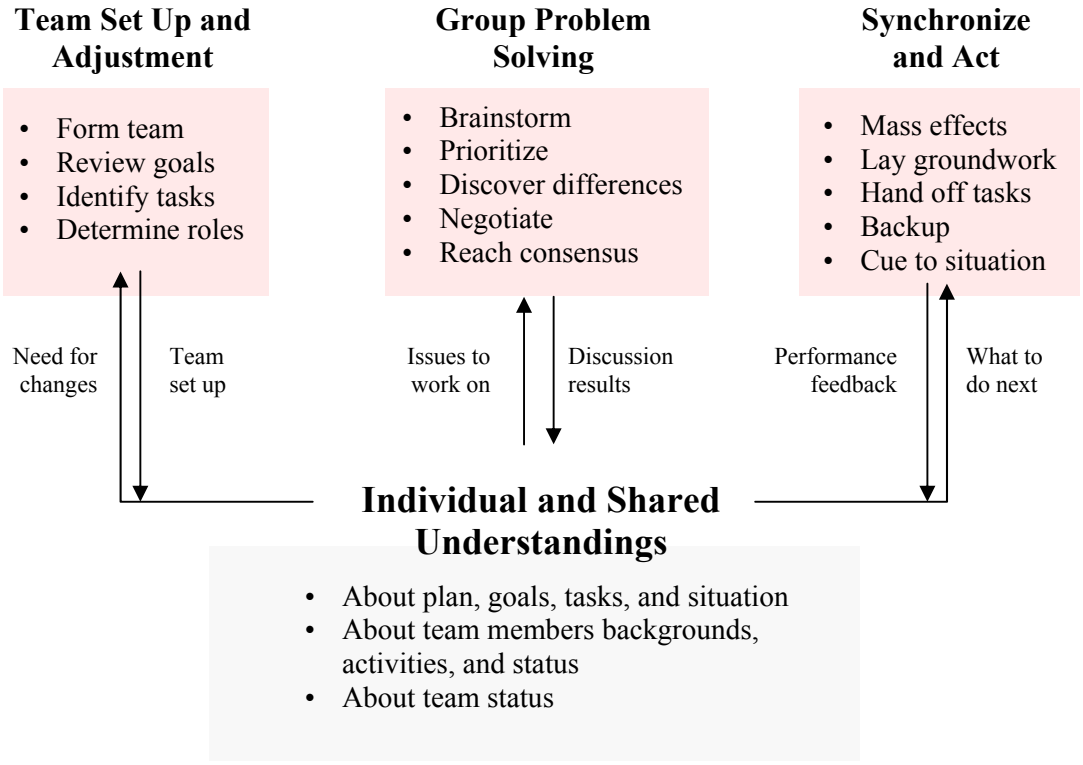

Figure 6. Building Blocks of Collaboration and Teamwork

The two way arrows in Figure 6 emphasize that the knowledge both enables and is enabled by the activities in the three upper boxes. Teams cannot carry out their tasks and work together effectively if they do not have the necessary knowledge. But because teams acquire the knowledge they need to do subsequent tasks by carrying out earlier tasks, they can't acquire the knowledge they need for future tasks if they fail in earlier ones. Thus, team failure can feed on itself, with early difficulties impeding task progress, which in turn impedes obtaining the knowledge required to continue working together in future tasks.

The audit trail documenting how infrastructure, process, or organizational changes impact team performance and products builds on the model of mutually reinforcing cycle of knowledge and behaviors depicted in Figure 6. Figure 7 "straightens" these iterative cycles to depict the causal chain for evaluating the impact of new information presentation and communication tools.

In this framework, all information presentations and communication have an impact initially by improving a person's knowledge or understanding. Thus, a new type of visualization that portrays uncertainty better will improve its user's understanding of uncertainty, which then leads to better job performance and teamwork.

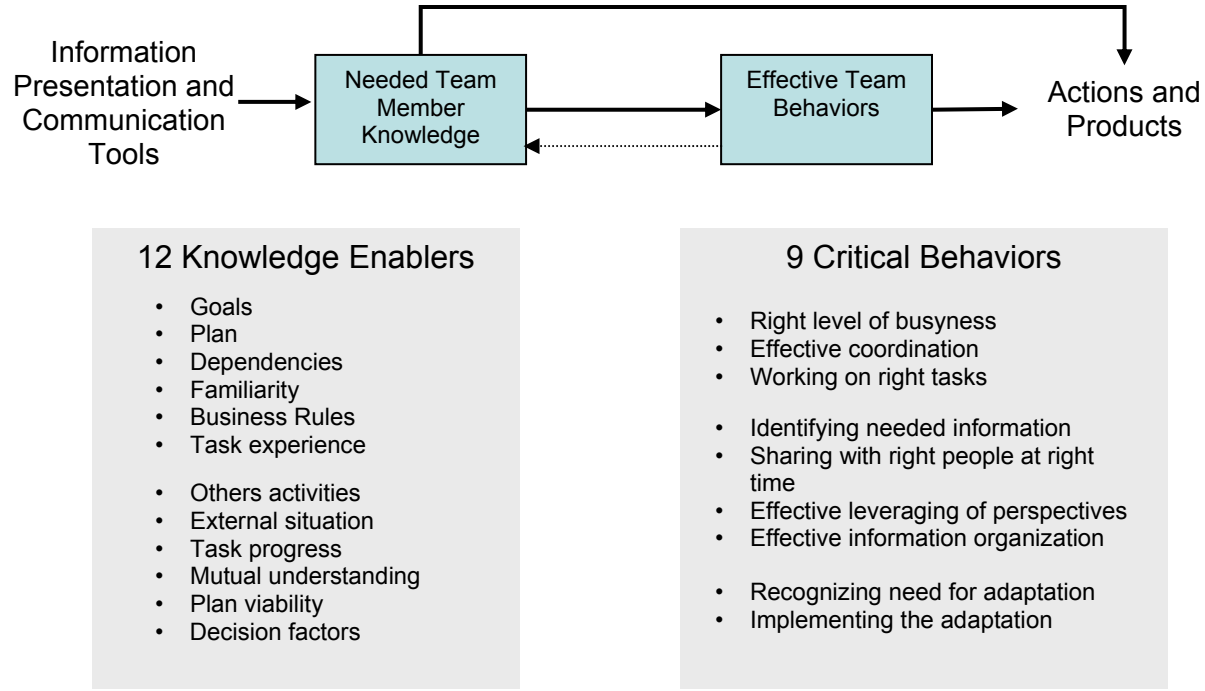

Figure 7. Enabling Knowledge and Behaviors

Our framework organizes the knowledge and understandings required for teams to work together into twelve categories. The first six of these categories encompass the knowledge obtained when a team organizes itself and learns to work together. This knowledge changes relatively slowly over time. The second six categories are the time sensitive understandings of team and task status and prospects at each instant of time. These understandings can change rapidly.

The audit trail framework organizes the critical team behaviors into nine categories. The first three of these concern how well the team coordinates and synchronizes its tasks. The next four categories concern how well the team manages and handles information. The last two categories address a team's ability to change when needed.

As described in the following section, the evaluation methodology provides methods for measuring each of the twelve enablers and nine critical behaviors. These measurements enable the evaluators to tell a compelling "story" about the reasons for a team's performance change. For example, a spatially distributed team may produce a product more efficiently when a tool that helps them be more aware of each others' activities is introduced. The overall performance metrics might show that the team is now creating a better product (as measured using the product metrics) faster and with fewer person hours. The behavioral metrics might then document that team members have reduced performing unnecessarily redundant tasks and members spend less time waiting for team members to finish precursor tasks. The knowledge metrics might document that team members are much more aware of what each other is doing, thus enabling the improved coordination. An analysis of the new information technology confirms that its displays are designed to help people know what others are working on.

This audit trail is important not only to document the reasons for increases effectiveness. It also provides an important safety net for evaluators to avoid premature and unwarranted conclusions when a new technology or process fails to lead to improvements.

A recent experiment at JFCOM was evaluating the advantages of an "Operational Net Assessment" (ONA), a key concept to support Effects Based Operations. In that experiment, the ONA was not shown to improve team performance.

Fortunately, the evaluators examined the experiment data to determine the reasons for this failure. They discovered that the improvements to team performance were blocked by problems with two of the knowledge enablers: goal understanding and business rules. Team members had different understandings of team goals, undermining their ability to work together. In addition, the newly formed team for the experiment had little experience working together, and had not established clear business rules for interacting with each other. Consequently, team members were unclear about when and how to share information and provide mutual support. These two factors accounted for the lack of observed team improvement. Being aware of these factor prevented the evaluators from incorrectly concluding the ONA could not be effective.

In later experiments, in which these limiting factors were overcome, ONA was shown to contribute to team effectiveness.

5. **Documenting the Evaluation Story: Data Collection**

Figure 8 portrays the elements of the causal audit trail (top sequence), and ties to each of these elements the data collection targets (bottom sequence) able to substantiate and document the audit trail story. The previous section provided a brief example of such a story: a tool improved awareness of team activities, which improved task coordination, which led to a better product. More generally, the story is that a tool, process, or organization improved critical knowledge in one or more of the twelve knowledge categories of Figure 7, which in turn improved the behaviors in some of the nine behavioral categories, which then improved team performance and product quality.

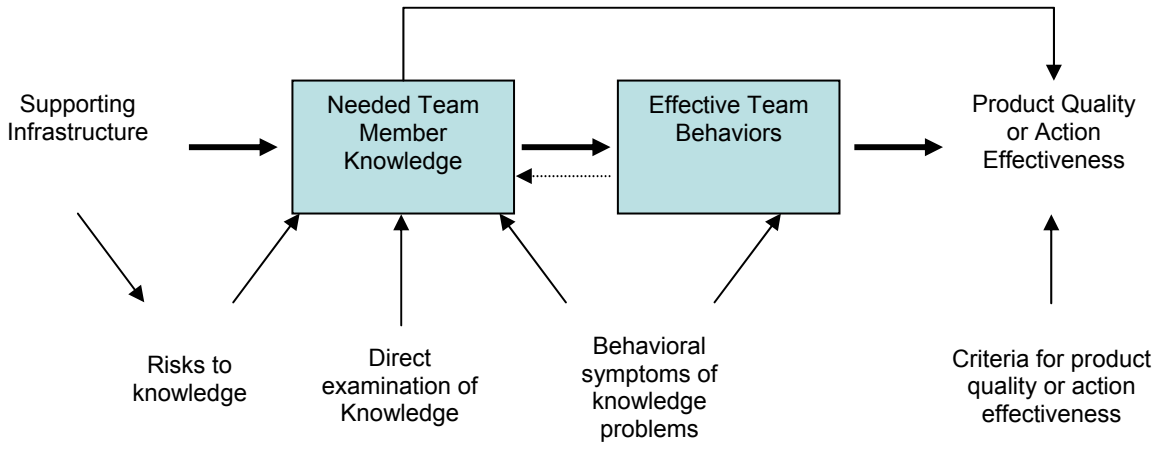

Figure 8. Data Collection Targets for Documenting the Evaluation Story

Documenting this story requires a theory to link tool, process, or organizational changes to knowledge, to link knowledge to behaviors, and to link knowledge and behaviors to improved product. Documenting the story also requires data to quantify changes in knowledge, behaviors, and product quality. The EBR handbook "Command Performance Assessment System" [Kirzl, 2003] describes the data collection and analysis to measure knowledge, behaviors, and product quality.

Risks to knowledge. As described in that handbook, the link between the supporting infrastructure (tools, processes, and organization) and knowledge are various risks to knowledge. These risks are task, team, and environmental factors that increase the difficulty of obtaining the knowledge needed for effective performance. Table 1 provides examples, selected from the more extensive set in the handbook, for these knowledge risks. The left column of the table shows illustrative tool services. The middle column lists knowledge risks that the tool service reduces. The right column is one or two of the knowledge enablers affected by that risk.

Tool and Tool Service	Knowledge / Understanding Risk	Key Knowledge Areas Impacted
Applications that enable team member's input (new material, comments) in near-real time	It is difficult to see other people do their jobs	Other's Activities
	It is difficult to link team products to the people who did them	Familiarity, Mutual Understanding
Monitors for watching others work	It is difficult to see other people do their jobs	Other's Activities
	Team members are sometimes assigned to tasks based on title rather than skill	Task Experience
Monitors focusing on external situation changes	It is a difficult environment in which to discover problems early	External Situation
	There are significant time lags between taking an action and knowing the result	External Situation, Decision Factors
	It is hard to see quickly the changes people make to either the situation or to team products	Other's Activities, External Situation

Table 1. Example of Handbook Table: Tools and Services that Reduce Knowledge Risks

Direct examination of knowledge. This is evaluation of the level of understanding in a knowledge area obtained by directly asking people questions about that knowledge area. Section 3.1 listed the questions for "understanding the external situation."(which can double as a product). The evaluation handbook lists example questions for measuring that and the other eleven knowledge areas. Example questions for "familiarity" (knowledge about others on team) are:

1. Who on the team are most knowledgeable about y?
2. Who has experience in subject y?
3. What is person z likely to think about y?
4. What is he most likely to do in situation y?
5. What are the conditions under which y is likely to need help with task z?

Behavioral symptoms. Observed behaviors contribute in two ways to documenting the impact of a tool, process, or organization on effectiveness. As symptoms of knowledge and understandings, they help document knowledge state. As observed behaviors, they document the extent to which the team's behaviors are effective in each of the nine behavioral categories listed in Figure 7.

Often, evaluators must interpret directly observed behaviors in order to link them either to the twelve knowledge or nine behavioral categories. Table 2 lists five symptoms extracted from a more comprehensive table in the evaluation handbook. The second column notes the data to be collected at each observed instance of a symptom. The third column scores how often the symptoms are observed, a count used to weight its significance.

Symptom	***Data to be Collected***	***Scoring***
People act in ways which the leader or sponsor believe are inconsistent with intent	Questionnaire or record leader feedback to staff	# of inconsistent actions per time period
Team members argue or disagree about what achievements constitute success	Record disagreements	# of disagreements/time period
Team members propose actions which if successful would be inconsistent with intent	Record actions. SME determine inconsistencies	Ratio of # of inconsistent actions to total actions
Sometimes team members pursue their own objectives rather than support team needs	Questionnaire	# of occurrences per time period
Team members state that some past team decision or orders contradicted overall intent	Questionnaire	# of occurrences

Table 2. Example of Handbook Table for Symptoms of Poor Goal Understanding

Each of the symptoms in the table can be a sign of poor understanding of goals. Unfortunately, most symptoms are ambiguous. The fourth symptom can also imply poor understanding of the plan or relationships. The fifth can imply poor understanding of decision factors.

Criteria for product quality of action effectiveness. Section 3 outlined the metrics for the three principal intellectual products: situation assessments, plans, and decisions. The handbook provides additional metrics for plans.

EBR has developed an additional set of metrics for action effectiveness. These metrics, used to support force transformation and network centric warfare, focus on team and organizational agility [Alberts, 2003]. These metrics for action effectiveness break out six components of agility. These are:

- Robustness, the ability to maintain effectiveness across a range of tasks, situations, and conditions
- Resilience, the ability to remain effective under attack or in a degraded state
- Responsiveness, the ability to react to a change in the environment in a timely manner

- Flexibility: the ability to identify multiple ways to succeed and the capacity to move seamlessly between them
- Innovativeness, the ability to do new things and to do old things in new ways
- Adaptiveness, the ability to change the work process, the ability to change the organization

6. **Summary**

Objective performance measures can greatly increase the credibility and utility of an evaluation, for they can measure the extent to which teams produce timely high quality products or act together effectively. Unlike subjective opinions of how well an individual or team is performing, objective measures always correlate with the extent to which assigned tasks are being carried out successfully.

Coupled with cause-effect models linking infrastructure to product, objective measures provide an audit trail of how a tool, process, or organizational change impacts performance and product quality. This audit trail includes documentation of the extent to which a tool, process, or organizational change has the properties able to facilitate acquisition of critical knowledge and understandings, objective measurement of these understandings, a documentation of the effectiveness of team behaviors resulting from this knowledge and understanding, and objective measures of team performance and product quality.

Useful models are now available that describe how individuals and teams create intellectual products. These models provide a theoretical basis for linking the infrastructure properties to knowledge, knowledge to behaviors, and knowledge and behaviors to team performance and product quality.

Data collection and analysis to support objective performance measurement have proved cost-effective. Data collection can be non-intrusive if necessary. Many of the evaluations conducted over the past twenty years did not require data collectors to directly question the evaluation participants.

The data collection and analyses methods to assess performance are too extensive to document fully in this paper. The EBR handbook "Command Performance Assessment System" describes extensively the data collection and analysis methods to measure knowledge, behaviors, and product quality.

7. **References**

Alberts, D.S., and R.E. Hayes. *Power to the Edge: Command… Control… in the Information Age*. Washington, DC: CCRP Publication Series. 2003.

Katzenbach, J.R. and D.K. Smith. "The Discipline of Teams." *Harvard Business Review*. 71 (2). Watertown, MA: Harvard Business School Publishing. 1993. pp111-120.

Kirzl, J, Noble, D., and Leedom, D. Command Performance Assessment System. Vienna, VA: Evidence Based Research. 2003.

Noble, D. *Enabling Effective Collaboration in Military Operations. Workshop Report.* Vienna, VA: Evidence Based Research. 2001.

Noble, D. "A Cognitive Description of Collaboration and Coordination to Help Teams Identify and Fix Problems." *Proceedings of the 7th International Command and Control Research Technology Symposium*. Quebec, Canada: Canadian Department of National Defense. 2002.

Noble, D. "Command Post of Future Decision Model." *Proceedings of the 2000 Command and Control Research Technology Symposium*. Monterey, CA: Navy Postgraduate School. 2002.

Wegner, D.M. "Transactive Memory: A Contemporary Analysis of Group Mind." Brian Mullen and George R. Goethals, eds. *Theories of Group Behavior.* New York, NY: Springer-Verlag. 1987. pp185-206.

Zsambok, C. E., G. Klein, M. Kyne, and D.W. Klinger. *Advanced Team Decision Making: A Model for High Performance Teams*. Fairborn, OH: Klein Associates, Inc. 1993.