2018

Feedback Design Considerations for Intelligent Team Tutoring Systems

Jamiahus Walton
Iowa State University, jwalton@iastate.edu

Alec Ostrander
Iowa State University, alecglen@iastate.edu

Kaitlyn M. Ouverson
Iowa State University, kmo@iastate.edu

Stephen B. Gilbert
Iowa State University, gilbert@iastate.edu

Michael C. Dorneich
Iowa State University, dorneich@iastate.edu

See next page for additional authors

Follow this and additional works at: https://lib.dr.iastate.edu/imse_conf

Part of the Operations Research, Systems Engineering and Industrial Engineering Commons

Recommended Citation
https://lib.dr.iastate.edu/imse_conf/131

This Conference Proceeding is brought to you for free and open access by the Industrial and Manufacturing Systems Engineering at Iowa State University Digital Repository. It has been accepted for inclusion in Industrial and Manufacturing Systems Engineering Conference Proceedings and Posters by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Feedback Design Considerations for Intelligent Team Tutoring Systems

Abstract
Challenges arise when developing a computer-based Intelligent Team Tutoring System (ITTS) that attempts to deliver feedback to teams as effectively as a human tutor. The purpose of this current work is to outline elements of feedback that should be considered when designing feedback for an ITTS. The authors present the results of a study that consisted of 32 participants grouped into 16 teams of two. Each team conducted a surveillance task where they received individual or team feedback. Feedback content was written using either the bald (direct feedback; no need for interpretation) or off-record (general feedback; interpretation needed) etiquette strategy. The results showed that feedback delivered using the bald etiquette strategy positively correlated with improved performance. The results also showed that team level feedback positively correlated with more accurate self-assessment among participants. This suggests that in an ITTS, direct feedback can lead to better performance, and that feedback provided at the team level can help to align self-interpretation of performance with actual task performance.

Disciplines
Operations Research, Systems Engineering and Industrial Engineering

Comments

Rights
Works produced by employees of the U.S. Government as part of their official duties are not copyrighted within the U.S. The content of this document is not copyrighted.

Authors
Jamiahus Walton, Alec Ostrander, Kaitlyn M. Ouverson, Stephen B. Gilbert, Michael C. Dorneich, Eliot H. Winer, and Anne Sinatra

This conference proceeding is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/imse_conf/131
Feedback Design Considerations for Intelligent Team Tutoring Systems

Jamiahus Walton¹, Alec Ostrander¹, Kaitlyn Ouverson¹, Stephen B. Gilbert¹, Michael Dorneich¹, Eliot Winer¹, Anne Sinatra²

¹Iowa State University, ²US Army Research Laboratory

Challenges arise when developing a computer-based Intelligent Team Tutoring System (ITTS) that attempts to deliver feedback to teams as effectively as a human tutor. The purpose of this current work is to outline elements of feedback that should be considered when designing feedback for an ITTS. The authors present the results of a study that consisted of 32 participants grouped into 16 teams of two. Each team conducted a surveillance task where they received individual or team feedback. Feedback content was written using either the bald (direct feedback; no need for interpretation) or off-record (general feedback; interpretation needed) etiquette strategy. The results showed that feedback delivered using the bald etiquette strategy positively correlated with improved performance. The results also showed that team level feedback positively correlated with more accurate self-assessment among participants. This suggests that in an ITTS, direct feedback can lead to better performance, and that feedback provided at the team level can help to align self-interpretation of performance with actual task performance.

INTRODUCTION

For years intelligent tutoring systems (ITSs) have been used to tutor individual in various fields, such as computer-assisted language learning, education, and military (Bradač & Kostolányová, 2016; Gamper & Knapp, 2002; Zachary et al., 1999). There is a growing interest in extending this technology to collaborative learning strategies such as team-based learning (Greenwald et al., 2017). Some principles of team-based learning include team formation, readiness assurance (individual and team), timely feedback, problem ordering, peer review, and incentive structures (Kibble, Bellew, Asmar, & Barkley, 2016). A human tutor can implement these principles and adapt to unexpected changes. Specifically, a human tutor is able to evaluate a team, develop appropriate feedback, and administer that feedback appropriately. However, challenges arise when attempting to develop an Intelligent Team Tutoring System (ITTS) that attempts to be as effective as a human tutor (Gilbert et al., 2018).

Team interactions are complex and difficult to measure. As a result, it is challenging for an ITTS to deliver feedback to a team in an effective manner that supports team development. The purpose of this work is to outline elements to consider when designing ITTS feedback. The next section discusses previous work that informs the design of ITTS feedback. The following section describes the preliminary results of a study that evaluated teams as they conducted a military surveillance task while receiving ITTS feedback during the sessions.

LITERATURE REVIEW

Feedback has been defined as a communication process that involves a source (i.e., an ITTS) sending a message to a recipient (Ilgen, Fisher, & Taylor, 1979). Feedback is a simple communication process but has complex consequences for the individual (Hattie & Timperley, 2007). Complexity grows when the focus includes both the team and individual (DeShon, Kozlowski, Schmidt, Milner, & Wiechmann, 2004).

There are many elements of feedback for teams that can be considered; to limit the scope of this work, the authors focus on peer monitoring, the content of feedback, the feedback modality, and the tutor architecture.

Peer Monitoring / Public Shaming

Peer monitoring takes place when corrective feedback is given to a single individual in the presence of a group. It is an accountability technique that has been used to discourage disruptive behaviors from coworkers or team members (Phillips, Zimmermann, & Bird, 2014). Peer monitoring can take many different forms, but two common types are indirect and direct. Direct peer monitoring has been defined as employees, or members of a group, monitoring coworkers and giving direct feedback that critiques their performance by providing praise, correcting, or reporting them. Indirect peer monitoring has been defined as employees who gossip and avoid co-workers who perform poorly (Loughry & Tosi, 2008). The authors focused on direct peer monitoring. Loughry and Tosi (2008) found that direct peer monitoring had a positive influence on performance when members had little supervision (i.e. low supervisory monitoring), but it had negative influence when there was high supervision. This suggests that a team can improve their performance by allowing them to manage and correct themselves. In the current study, low supervision was provided to determine if this applies in an ITTS.

Feedback Language, Content, and Affect

The purpose of an ITS is to effectively communicate the knowledge within the system (Shute & Psotka, 1996). During human-human interaction, there are social etiquette rules that dictate how humans interact with one another. A question that arises is whether computers, when it comes to human to computer interaction, should be guided by the same etiquette rules humans adhere to, or should there be a different set of rules (Meyer, Miller, Hancock, de Visser, & Dorneich, 2016).

A reasonable question is whether applying conversation etiquette rules to feedback has any effect on behavior. Brown and Levinson (1987) describe several etiquette strategies used
during human to human interaction. Yang and Dorneich (2016) found that manipulating etiquette strategies when delivering feedback can have a substantial impact on behavior, as well as motivation and confidence. To limit the scope of this current work, the authors focus on direct peer monitoring and off-record and bald etiquette. Off-record etiquette is characterized by general or indirect feedback such that the receiver must interpret the tutor’s intention. For example, a tutor may say, “It’s important to confirm at appropriate times.” The person who receives this feedback must interpret why they received that feedback. In contrast, bald etiquette is direct, given in a way that little to no interpretation is needed. An equivalent bald statement could be, “You need to work on confirming at appropriate times.” The authors sought to understand how bald and off-record etiquette influence performance by evaluating an ITTS that uses these strategies.

More work is needed to offer insight on how different elements of feedback influence team performance. To limit the scope of this current work, the authors focus on the effects of feedback target (i.e., individual vs. team) and etiquette strategy (i.e., bald vs. off-record). The authors evaluated an ITTS that delivered feedback focused on either the team or individuals and employed bald and off-record etiquette.

Tutor Architecture

An ITTS capable of effectively providing feedback at either the individual or team level as described above requires careful design of the tutor architecture. While single-user ITSS are common, the challenge of team training had not been met (Bonner, Walton, Dorneich, Gilbert, & Sottilare, 2015). The authors extended the Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare, Brawner, Goldberg, & Holden, 2012) to support teams (Gilbert et al., 2018).

The first complication comes in assessment. An individual tutor may be complex in its assessment methods, but it only ever needs to assess one user. In contrast, a team tutor must assess multiple users simultaneously. An ITTS must model each individual’s behavior, as well as the interactions of the team and each subgroup of members. Because the number of required models grows factorially with the number of learners on the team, this quickly becomes a difficult problem for large teams. As the system developed by the authors is meant to be a proof of concept, the scenario used in this study was a simple, two-person team task.

Once a tutor has accounted for user assessments, it must decide whether feedback is needed. For instance, should the team tutor decide to give feedback based on the performance of each individual user, when the team shows emergent behaviors, or somewhere between? Should feedback be delivered only to the user whose performance triggered the feedback, or to everyone on the team? In human training situations, it has been shown that receiving others’ feedback may encourage peer monitoring behavior, but it may also be perceived as unneeded or distracting (Loughry & Tosi, 2008). However, it remains to be seen whether an ITTS can elicit the same effects. The system implemented by the authors can be set to use either strategy, and the study in the following section explores these strategies.

Hypotheses

The hypotheses of this study were as follows:

1. Participants who receive team feedback will perform better than teams that received none or individual feedback due to effects of peer monitoring.
2. Bald feedback will lead to more improved performance than off-record feedback.

METHODS

Participants

This study had 32 participants (22 male, 10 female), aged 18-35, grouped into 16 teams of 2. Additional teams were also recruited, but their data were lost due to technical issues.

Surveillance Task

The ITTS was designed to accompany a military team surveillance training task, referred to in this paper as the Surveillance Task. The task was designed to be a conceptually simple yet reusable testbed for ITTS evaluation (Bonner et al., 2015). It was developed in the Virtual Battle Space 2 (VBS2) computer game engine and placed two users overlooking a 3D desert environment on opposite sides of a building roof. VBS2 was integrated with the GIFT software, and assessment as well as feedback occurred in GIFT during the Surveillance Task. The team members were each responsible for monitoring half of the environment surrounding them (their zones) and scanning for mobile opposing forces (OPFOR) (see Figure 1).

![Figure 1: Top view of the surveillance scenario with running OPFOR as red diamonds](image)

When a user observes an OPFOR heading out of their zone towards their team mate’s zone, they should make the team mate aware of the incoming target and on which side it is crossing (using the green poles as a reference). The second user should then acknowledge having heard the first user’s communication and confirm the target when they see it in their zone. For evaluation, these three subtasks were formalized as Transfer, Acknowledge, and Identify. The following section provides more detail on assessment methods.
Participants were trained and expected to press keyboard keys corresponding to the 3 types of communication so that the system could receive team communication input in real time. Each of these communications was recorded by the ITTS in conjunction with the environment state to re-evaluate each user model in real time. When the ITTS found that feedback was needed, it was presented to the user visually on a panel aligned to the left of the VBS2 scenario. Figure 2 shows an example of the environment and tutor interface.

Feedback Implementation

Feedback target indicates whether assessment and feedback should occur at the team or individual level. For example, a certain statement may be set to display if three Transfers actions have been missed recently. With an individual feedback target, the feedback is triggered once a single user has made that mistake three times and will display only to that user. In contrast, with a team feedback target, the feedback is triggered once three Transfers are missed between the two users and displayed to both simultaneously.

Etiquette strategy refers to the precise language being used in the content of the feedback. The authors focused on two specific strategies: bald and off-record etiquette. Feedback statements employing each of these strategies were interspersed in the options available to the ITTS.

Table 1: ITTS feedback options for Individual-level Teams

<table>
<thead>
<tr>
<th>Question</th>
<th>Assess.</th>
<th>Statement</th>
<th>Etq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans.?</td>
<td>Below</td>
<td>It is important to communicate crossings</td>
<td>O.R.</td>
</tr>
<tr>
<td>Trans.?</td>
<td>Below</td>
<td>Report transferring OPFOR to team by pressing 1 or 2 key</td>
<td>Bald</td>
</tr>
<tr>
<td>Trans.?</td>
<td>Below</td>
<td>It is important to communicate crossings. Your communication needs work</td>
<td>O.R.</td>
</tr>
<tr>
<td>Tr. Time?</td>
<td>Above</td>
<td>Successful handoff</td>
<td>Bald</td>
</tr>
<tr>
<td>Tr. Time?</td>
<td>At</td>
<td>Make sure you are not transferring too early</td>
<td>Bald</td>
</tr>
<tr>
<td>Tr. Time?</td>
<td>Below</td>
<td>It is important to communicate when an OPFOR crosses into your partner’s zone</td>
<td>O.R.</td>
</tr>
<tr>
<td>Ack. Time?</td>
<td>Above</td>
<td>Successful confirmation</td>
<td>Bald</td>
</tr>
<tr>
<td>Ack. Time?</td>
<td>At</td>
<td>Acknowledge your communications as soon as you receive them</td>
<td>Bald</td>
</tr>
<tr>
<td>Ack. Time?</td>
<td>Below</td>
<td>It is important to confirm at appropriate times</td>
<td>O.R.</td>
</tr>
<tr>
<td>ID Time?</td>
<td>Above</td>
<td>Excellent work identifying OPFOR</td>
<td>Bald</td>
</tr>
<tr>
<td>ID Time?</td>
<td>At</td>
<td>It’s important to identify OPFOR as quickly as possible</td>
<td>O.R.</td>
</tr>
<tr>
<td>ID Time?</td>
<td>Below</td>
<td>Alpha Team, Identify OPFOR immediately</td>
<td>Bald</td>
</tr>
</tbody>
</table>

Feedback for each assessment level was designed for four separate questions: 1) Did Transfer occur? 2) Was Transfer timing ok? 3) Was Acknowledge timing ok? and 4) Was Identify timing ideal? Teams or individuals received the feedback in Table 1 corresponding to their performance on each question. Most questions had corresponding feedback statements for performance assessed as Above Expectation, At Expectation, or Below Expectation.

Variables

The study had two independent variables: trial (within-subjects) and feedback target (between-subjects). Trials were not identical since OPFOR movement was semi-random, but otherwise there were no distinct differences between trials. Teams were randomly assigned a feedback target, with 5 teams receiving Individual feedback, 5 teams receiving Team feedback, and 6 teams receiving no feedback. The etiquette strategies used by the ITTS are also analyzed as a within-subjects variable in this paper; however, no manipulation of etiquette was performed since the specific feedback statements presented depended on the participants’ performances.

There were several dependent variables. The authors synthesized information for every action participants took during the task. As a preliminary metric, the number of each type of error – Transfer presence and timing, Acknowledge timing, and Identify timing – was used as a measure of task performance. These were divided into “miss” type errors – when the participant did not perform an action they should have, and “extra” type errors – when the participant performed an action they didn’t need to do. There are also subjective data from surveys given to participants between trials measuring self-assessment, team assessment, taskload, and feedback perception. This paper analyzes participants’ self-assessment of task performance accuracy. This was measured after every trial on a scale of 0-100.

Procedure

Participant teams were randomly assigned to receive either team feedback, individual feedback, or no feedback (control group) and entered separate rooms for the study. Each room had a computer in it. During the scenario, all team communication occurred over an open audio channel. After receiving video instructions and completing a training session on the surveillance task, participants partook in four five-minute trials. After each trial, participants were asked to complete a short post-trial survey. At the end of the study, the participants also completed a post-experiment survey and participated in a debriefing session with the experimenters.

RESULTS

The main purpose of this study was to demonstrate and evaluate the efficacy of the ITTS. Initial data on learners’ overall performance in the ITTS are reported in Bonner et al. (2017). This paper, however, analyzes the specific impact of the feedback design in the Surveillance Team Tutor. The analyses below explore the effects of different feedback strategies on participant performance and self-assessment.
**Task Performance**

A series of two-way mixed ANOVAs was performed to determine if there was an effect of feedback target (between-subjects) and trial (within-subjects) on the number of any of the types of errors made and if there was an interaction effect. Because the data did not meet several assumptions of the ANOVA procedure, the Aligned Rank Transform was performed on all data (Wobbrock, Findlater, Gergle, & Higgins, 2011). There were no outliers in the transformed data, as assessed by inspection of studentized residuals. Ranks were evenly distributed, as assessed by normal Q-Q plot. There was homogeneity of covariances, as assessed by Box’s M test. Not all groups exhibited homogeneity of variance and sphericity, assessed by Levene’s test and Mauchly’s test, respectively. In these cases, the analysis was completed using the Greenhouse-Geisser correction. Non-parametric tests (Friedman Test and Kruskal-Wallis H Test for trial and target, respectively) were run to back up and verify the results.

No statistically significant interaction effect between trial and feedback target on any of the error counts was found. The main effect of trial was statistically significant on all “miss” type errors, $F(3, 87) = 7.128, p = .002$, partial $\eta^2 = .197$. Post hoc analyses found that Trial 3 had statistically significantly fewer errors than Trials 1, 2, and 4. Trials 1, 2, and 4 were not statistically significantly different in terms of errors. There was no statistically significant main effect of feedback target in any of the errors counts.

To evaluate if there was a relationship between the feedback etiquette used and performance, correlations were performed between the number of each type of feedback received and the number of each type of error the participant made. Shapiro-Wilk’s test indicated that not all variables were normally distributed, so Spearman’s rank-order correlation was chosen. Monotonic relationships for each correlation were confirmed by visual inspection of scatterplots.

The results of correlations between the type of feedback etiquette used and participant performance are presented in Table 2 below, with statistically significant correlations starred *. In this context, a negative correlation is desired since well-designed feedback should lead to fewer errors. It might be hypothesized that fewer errors led to more feedback; this was not the case, however, based on how the software was designed. Overall, more bald etiquette feedback is highly correlated with fewer errors made, while off-record feedback has no statistically significant relationship. This trend could also be seen in scatterplots within each trial; however, the sample sizes were too small to verify statistical significance.

**Table 2: Correlations between number of bald and off-record feedbacks received and number of errors made across all participants, all trials (n=128)**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Bald Feedback</th>
<th>Off-Record Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer</td>
<td>-.283 (p = .001) *</td>
<td>-.014 (p=.879)</td>
</tr>
<tr>
<td>Acknowledge</td>
<td>-.286 (p = .001) *</td>
<td>-.071 (p=.425)</td>
</tr>
<tr>
<td>Identify</td>
<td>-.072 (p = .419)</td>
<td>.039 (p=.658)</td>
</tr>
<tr>
<td>Total Extras</td>
<td>-.251 (p = .004) *</td>
<td>-.104 (p=.242)</td>
</tr>
<tr>
<td>Total Misses</td>
<td>-.255 (p = .004) *</td>
<td>.015 (p=.864)</td>
</tr>
<tr>
<td>Total Errors</td>
<td>-.299 (p = .001) *</td>
<td>-.037 (p=.678)</td>
</tr>
</tbody>
</table>

**Self-Assessment**

In addition to improving actual performance, another important goal of feedback is to make learners aware of their own performance level and where they need to improve. Because the self-assessment measure (subjective 0-100 rating) is not on the same scale as the performance measure (total number of errors), a direct comparison without bias is not possible. To measure this effect, correlations between number of errors and self-assessment were run for each feedback target group. Monotonic relationships for each correlation were confirmed by visual inspection of scatterplots (Figure 3). Data were positively-skewed as assessed by boxplot, but Spearman’s correlation yielded results similar to Pearson’s correlation, indicating the analysis was not affected.

![Figure 3: Scatterplots of self-assessment over number of errors made, by feedback target group. Each point represents one participant in one trial. Counts do not exactly match participant counts due to lost survey data.](image)

The No Feedback group had self-assessment ratings moderately correlating with errors made, $r = -.31, p = .049$. 
The Individual Feedback group did not yield a statistically significant correlation, \( r = -0.174, p = .415 \). The Team Feedback group had self-assessment ratings strongly correlating with errors made, \( r = -0.76, p < 0.001 \).

The Fisher Z-Transformation was used to evaluate statistical differences between these correlations. The Team Feedback group (\( z' = .997 \)) was found to have statistically significantly stronger performance/self-assessment correlations than both the No Feedback group (\( z' = .328, p = .007 \)) and the Individual Feedback group (\( z' = .339, p = .022 \)). The strengths of correlation of the latter two groups were not statistically significantly different.

**DISCUSSION**

The goal of this work was to explore elements that should be considered when designing feedback for use in an ITTS. Specifically, we evaluated the effects of feedback target and etiquette strategies.

The results of this study support hypothesis 1, that ITTS users were able to use bald feedback to improve their performance much more effectively than off-record feedback. This is expected for real-time feedback in a high workload scenario; users may focus on the task and ignore feedback that is not immediately deemed helpful (von Muhlenen, Rempel, & Enns, 2005). Although hypothesis 2 was not supported, teams with team-level feedback did demonstrate stronger self-assessment of their performance level than their peers. This result is surprising, as one would expect individual feedback to result in better individual self-assessment and team feedback to result in better team assessment. More work is needed to validate and explain this result.

It should be noted that the possibility of these results being skewed by practice effects over the four trials could not be entirely ruled out due to the small sample size. With that said, all precautions were taken to ensure their validity.

This paper presents the preliminary findings of a larger study assessing the first ITTS developed using GIFT. There are still many questions in this nascent area of research. As the first team GIFT study, a simple team of 2 members with identical roles was used. Future studies will increase the number of team members and diversify the team structure.

The application of etiquette strategies to ITTS feedback has shown promise. Further studies detailing this effect for several strategies and scenarios will be invaluable to the field. Future studies should also consider the effect of indirect peer monitoring (Loughry & Tosi, 2008).

More work is needed to investigate the influence of other feedback elements, such as focus level, modality, and timing.

**CONCLUSIONS**

The development of future ITTSs will depend on the support of powerful system architectures, but the success of these systems for learning and training will depend just as heavily on the intelligent design of feedback the systems will give to users. This paper discusses several of the concepts that must be considered to design effective feedback for an ITTS and demonstrates the effects of using two of these concepts to design feedback for a prototype ITTS.

**REFERENCES**


