Objective. This study investigated adapting the interaction style of intelligent tutoring system (ITS) feedback based on human-automation etiquette strategies.

Background. Most ITSs adapt the content difficulty level, feedback timing, or provide extra content when it detects cognitive or affective decrements. Our previous work demonstrated that changing the interaction style via different feedback etiquette strategies have differential effects on students’ motivation, confidence, satisfaction, and performance. The best etiquette strategy was also determined by user frustration.

Method. Based on these findings, a rule set was developed that systemically selected the proper etiquette strategy to address one of four learning factors (motivation, confidence, satisfaction, and performance) under two different levels of user frustration. An experiment explored whether etiquette strategy selection based on this rule set (systematic) or random changes in etiquette strategy for a given level of frustration impacted the four learning factors. Participants solved mathematics problems under different frustration conditions with feedback that adapted dynamic changes in etiquette strategies either systematically or randomly.

Results. The results demonstrated that feedback with etiquette strategies chosen systematically via the rule set could selectively target and improve motivation, confidence, satisfaction, and performance more than changing etiquette strategies randomly. The systematic adaptation was effective no matter the level of frustration for the participant.

Conclusion. If computer tutors can vary the interaction style to effectively mitigate negative emotions, then ITS designers would have one more mechanism in which to design affect-aware adaptations that provide the proper responses in situations where human emotions affect the ability to learn.

Keywords: Adaptive Automation, Affective Factors, Intelligent Tutors, Human-Computer Interaction, Etiquette
INTRODUCTION

Cognition and emotion (affect) are complementary processes that enable student learning (Lehman, D’Mello, & Person, 2010). Positive emotions can result in improved information processing decision making, creative problem solving, cognitive flexibility, and memory retention (Contai & Manske, 2009; Erez & Isen, 2002; Isen, Daubman, & Nowicki, 1987; Isen, Shalker, Clark, & Karp, 1978). Negative emotions, notably frustration, can lower productivity, increase decision-making time, and lower learning efficiency (Klein, Moon, & Picard, 2002; Powers, Rauh, Henning, Buck, & West, 2011; Lerner, Li, Valdesolo, & Kassam, 2015; Graesser, Chipman, Haynes, & Olney, 2005).

Tutors are keenly aware of the importance of emotions, and actively use strategies to influence human emotions to improve student learning (Petrovica, 2013). Tutors may avoid disagreement to maintain non-adversarial relationships, or remain optimistic to encourage students when they struggle (Pearson, Kreuz, Zwaan, & Graesser, 1995). Tutors spend as much time attending to student emotional needs as cognitive learning needs (Lepper & Chabay, 1988). Intelligent tutoring systems (ITSs) are computer-based instructional systems that provide one-to-one feedback to enable learning by modifying instructional content, timing, and teaching strategies (Wenger, 1987; Murray, 2003; Koedinger & Tanner, 2013; Gilbert, Blessing, & Guo, 2015). ITS design has more frequently focused on cognitive aspects of learning, such as assessing student content knowledge to trigger tutor feedback (Wood & Wood, 1999; Zakharov, Mitrovic, & Johnston, 2008; Roll, Aleven, McLaren, & Koedinger, 2011). This contrasts with human tutors, who adapt behavior to meet the emotional needs of students (Woolf et al., 2009).

Evaluating learning effectiveness using only assessments of performance ignores other factors influencing learning. Keller (2009) identified multiple learning factors for sustaining student motivation and performance: Attention, Relevance, Confidence, and Satisfaction (ARCS). Higher levels of these factors correlate with higher levels of engagement with the material (Mohammad & Job, 2012) and increased learning effectiveness (Keller, 2009). ITSs have not always achieved the same impact as a good human tutor (Fletcher-Flinn & Gravatt, 1995). The ability to incorporate affect into ITS systems may partially address this gap (Zakharov et al., 2008). Researchers have developed affect-aware systems with the ability to detect and adapt to user emotional states (Picard, 1997; D’Mello et al., 2008; Woolf et al., 2009; Wang & Johnson, 2008; Grawemeyer et al., 2015).
Much work has focused on the detection of emotion. For instance, emotion has been shown to be related to mental workload and mental engagement, two constructs that can be detected through EEG-based indexes (Chaouachi, & Frasson, 2012). One barrier to wide application has been the difficulty in accurately classifying emotional states in a low-cost, unobtrusive manner.

Increasingly researchers are also studying how to prevent negative emotional states (Zakharov et al., 2008), regulate emotions (Malekzadeh, Mustafa, & Lahsasna, 2015), and mitigate the effects of emotions (Petrovica, 2013; Jraidi, Chalfoun, & Frasson, 2012; D’Mello, & Calvo, 2011). While some emotions (delight, boredom) are susceptible to the appropriateness of the feedback, negative emotions such as frustration have been more robust, where any attempt to alleviate frustration usually results in some benefit to the student (Robison, McQuiggan, & Lester, 2009). Theoretical frameworks can inform intervention design. For instance, self-regulation models (Boekaerts & Corno, 2005) address the ability of students to mitigate emotions to stay in the zone of proximal development; neither too frustrated or too bored (Murray & Arroyo, 2002). Negative emotions can decrease self-regulation, leading to poorer learning outcomes, with the opposite effect for positive emotion (Pekrun, Goetz, Titz, & Perry, 2002). Recent work has focused on the role confusion plays in benefitting learning (D’Mello, Lehman, Pekrun, & Graesser, 2014). Confusion during the learning process can be beneficial if students work to overcome the source of misunderstandings, if they remain engaged and motivated (Robison, McQuiggen, and Lester, 2009).

Adaptive ITSs automatically tailor behavior in real-time to best support learning. Adaptive systems have four general categories of modification: function allocation, task scheduling, interaction style, and content (Feigh, Dorneich, & Hayes, 2012). Adaptations to negative emotions have typically been task-focused or emotion-focused actions (Malekzadeh et al., 2015). Task-related strategies include changing task difficulty (Harley, Lajoie, Frasson, & Hall, 2015), adjusting timing and difficulty of assessments (Arroyo et al., 2014), or providing additional examples and hints (Chaffar, Derbali, & Frasson, 2009; Woolfe et al, 2009). Emotion-focused actions include providing empathetic responses (D’Mello, & Graesser, 2012; Mao & Li, 2009), mirroring the learner emotions (Picard et al., 2004; Zakharov et al, 2008), or providing behavioral prompts (D’Mello, & Graesser, 2015).

What has been less explored is modification of the interaction. In this work, changes to the interaction style of an ITS are produced by changes in the politeness level of a feedback
statement, but not its amount of information or subject matter. Our work is built upon the premise, established by previous work, that intelligent tutoring systems should employ social intelligence, where *how* a tutor provides feedback is as important as the cognitive intelligence of *when* and *what* feedback is presented (McLaren, DeLeeuw, & Mayer, 2011; Lepper & Woolverton, 2002). This work explores the viability of changing the interaction style of ITSs when providing feedback, without changing content difficulty level, feedback timing, or feedback information content. The concept of etiquette strategies may provide design inspiration on how to adapt interaction style.

Social interactions between humans are governed by expectations based on conventional norms. Etiquette is a code of requirements for social behaviors (Brown & Levinson, 1978). The linguistic aspects of etiquette of human-human interaction have identified independent factors in politeness including social power (e.g., relative positions in social hierarchies, age, gender), social distance (i.e., politeness increases with familiars, but decreases with both intimates and strangers), and imposition (e.g., requesting, urgency, apologizing, thanking, indebtedness, complaining). For a review, see Kasper (2005). Social power and distance are decided by the relationship between speaker and hearer. Level of imposition refers to the amount of demand or burden on the hearer. While the social power and the social distance between two people only change slowly over time, the imposition from speaker to hearer can be adjusted moment-to-moment.

Brown and Levinson (1987) identified four types of etiquette strategies to protect an individual’s self-esteem (“face threat”) from imposition: bald, negative politeness, positive politeness, and off-record (for examples, see Table 1). In bald strategy, the speaker speaks in a direct way without consideration of the imposition level on the hearer, or minimizing threats to the hearer's face. Positive politeness minimizes social distance between speaker and hearer by expressing statements of friendship, solidarity, and compliments. Negative politeness attempts to be respectful and assumes imposition on the hearer. Off-record gives indirect feedback, requiring the hearer to infer what was intended. Adaptation of the style of interaction using different etiquette strategies would alter the level of face threat which the imposition causes in communication. Pearson et al. (1995) demonstrated that human tutors varied between three different etiquette strategies when interacting with students to enhance or inhibit tutoring: bald, positive politeness, negative politeness.
Table 1. Example sentences of etiquette strategies for feedback regarding formulas (example 1) and drawing (example 2).

<table>
<thead>
<tr>
<th>Etiquette Strategies</th>
<th>Example 1 sentences (subject is formulas)</th>
<th>Example 2 sentences (subject is drawing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bald</td>
<td>Use appropriate formula.</td>
<td>Draw the problem out to understand it better.</td>
</tr>
<tr>
<td>Positive Politeness</td>
<td>Why don't you try other formulas? Let's check them together!</td>
<td>Let's try to draw a picture to better understand what the problem is asking.</td>
</tr>
<tr>
<td>Negative Politeness</td>
<td>If it's alright with you, could you please check other formulas as well?</td>
<td>I don't mean to interrupt you but it would help if you draw out the problem.</td>
</tr>
<tr>
<td>Off-Record</td>
<td>Various formulas are provided.</td>
<td>I wonder what this problem looks like visually.</td>
</tr>
</tbody>
</table>

This work investigates the premise that human-human etiquette strategies are generalizable to human-computer etiquette. Users tend to personify their computers by attributing human-like characteristics to them, and react to computers based on expectations from human-human interaction (Nass, Steuer, & Tauber, 1994). The concept of human-human etiquette has been applied in human-computer interaction (Miller, Wu, & Funk, 2008; Hayes & Miller, 2010; Miller, Wu, & Ott, 2012; Dorneich, Ververs, Mathan, Whitlow, & Hayes, 2012). Studies have operationalized politeness into various types of etiquette strategies. Johnson and Rizzo (2004) developed a model of politeness in tutor dialog that incorporated politeness factors of power and distance, and learner states of motivation and confidence, to automatically generate natural language dialogs for use in a foreign language tutor that generated text-based feedback in one of multiple politeness strategies: exaggerate, common ground, be vague, understate, question, tautology, impersonalize, and indirect. McLaren et al (2007) operationalized politeness by rephrasing all feedback to be more “polite” and simultaneously improving both its positive and negative face. In the current study, we operationalized Brown and Levinson’s (1987) etiquette strategies and focused on changes to the redress of imposition of feedback.

In preliminary work, different etiquette strategies and levels of frustration combined to have different impacts on the learning factors of performance, motivation, confidence, and satisfaction (Yang & Dorneich, 2016). For instance, bald etiquette strategy led to higher performance, positive politeness to higher subjective ratings of motivation and satisfaction, and negative politeness to higher subjective ratings of confidence and satisfaction. However, the most effective strategy for a given learning factor sometimes changed with frustration. The
current study builds upon these results to develop a rule set (see next section) that specifies under which conditions (triggers) different strategies should be used to provide feedback during students’ problem-solving. It is an open question whether a generalized set of rules can be effective for a population of users, and whether they can be designed in such a way to provide one-on-one interaction customized to the individual. Computer tutors more responsive to their students emotions may help students persist in their learning even when frustrated or motivation might otherwise flag.

Studies comparing polite feedback with direct feedback have mixed results. Wang and Johnson (2008) found evidence of learning gains for some types of questions but not others, and gains in some measures of motivation. In evaluations of web-based tutoring, McLaren et al. (2011) found lower prior-knowledge students benefited more from polite than direct feedback, although the overall effect was weaker in the classroom than the lab. In this current study, our aim was to test if differences within polite feedback can be used to target and improve the learning factors of motivation, confidence, satisfaction, and performance while mitigating the effect of frustration.

**ADAPTIVE TUTORING SYSTEM PROTOTYPE**

An adaptive tutoring prototype interface developed for this evaluation provided college-level algebra, geometry, trigonometry, calculus, statistics, and probability problems (Figure 1). Problems could be labeled as “easy” or “hard” to describe their level of difficulty. Research has shown that the level of frustration people experience can be affected by how confident they are in their ability to achieve the goal (Bessiere, Ceaparu, Lazar, Robinson, & Shneiderman, 2004). When a learner expects a problem to be easy but then finds it difficult, a learner’s expectations of goal attainment and satisfaction are unfulfilled, which can lead to frustration (Berkowitz, 1989). Based on the results of a pilot study, labeling hard question as easy was used in the evaluation to manipulate frustration. In addition, the evaluation verified that this manipulation did significantly change perceived frustration in participants.
The system provided recorded voice feedback while students solved problems. Voice and appearance can give cues to factors such as gender, age, education level, culture, and socio-economic position. Therefore recordings were designed to be homogenous and neutral, using the same voice, intonation, accent, and speaking pace. In a Wizard-of-Oz evaluation (Dahlbäck, Jönsson & Ahrenberg, 1993), while the participant solved math problems, the experimenter observed progress and triggered appropriate feedback voice files, based on a rubric (Table 2) for specific problem-solving errors or misconceptions (Gordon, 2008).

Table 2. Errors or Misconceptions that Trigger Feedback for different problem-solving steps.

<table>
<thead>
<tr>
<th>Step</th>
<th>Problem Solving Step</th>
<th>Errors or Misconceptions that Trigger Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identify and define the problem or situation</td>
<td>• Does not know how to start</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Does not know which variables are defined</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Does not know how to define variables</td>
</tr>
<tr>
<td>2</td>
<td>Generate alternatives</td>
<td>• Does not know which variables to solve for</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cannot find the equation to use</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Does not know how to approach</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Miss other variables</td>
</tr>
<tr>
<td>3</td>
<td>Evaluate the alternative suggestions</td>
<td>• Is not sure this approach is correct</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Does not solve all variables</td>
</tr>
<tr>
<td>4</td>
<td>Make the decision</td>
<td>• Cannot decide on the calculation methods</td>
</tr>
<tr>
<td>5</td>
<td>Implement the solution or decision</td>
<td>• Does not realize they are finished</td>
</tr>
</tbody>
</table>

Each step had at least one proactive (e.g., “Define the variables”) and at least one reactive (e.g., “It’s not the appropriate formula”) feedback comment. Feedback was specific to each math problem and differed between problems. Four versions of each feedback, one in each etiquette strategy, were designed to be homogeneous in content by restricting each feedback utterance to one subject (e.g., formula, drawing – Table 1), and varying the politeness vocabulary. Therefore,
length of utterances differed. Future work could explore if feedback length has an impact beyond the type of politeness used.

The adaptive tutoring system was designed to improve four learning factors associated with the ARCS model: motivation, confidence, satisfaction, and performance. A rule set was developed to trigger the most appropriate etiquette strategy as the basis for systematic adaptation, by targeting the learning factor most in need of improvement and applying an etiquette strategy to improve it. The process is outlined in Figure 2.

![Figure 2. The logical flow of the tutor to decide which learning factor to target, and which etiquette strategy to use when providing feedback in the next problem.](image)

After each math problem, the student rated motivation, confidence, and satisfaction on a 10-point scale, and a performance score was calculated (step 2). An algorithm determined which learning factor should be “targeted” next (step 3). First, if any learning factors’ ratings <= 5, target factor with the lowest score. If multiple factors shared the same lowest rating, the ARCS model determined priority order (highest to lowest): motivation, confidence, satisfaction, performance. If all ratings > 5, choose factor with the largest decrease from the previous measurement. Ties were broken using the same order as above.

Once the targeted learning factor (step 3) and the level of frustration (step 1) were identified, the etiquette strategy was chosen (step 4) from options listed in Table 3. Previous work established an initial, course-grained guidance on the most effective etiquette strategy given a learning factor and frustration level (Yang & Dorneich, 2016). Feedback during the next problem was delivered using the chosen etiquette strategy (step 5). Some combinations had more than one etiquette strategy identified as effective. If a strategy failed to improve the targeted
factor, then an alternative strategy was chosen the next time that factor was targeted. This enables some individual customization as the system learns which strategy is most effective.

<table>
<thead>
<tr>
<th>Learning Factor</th>
<th>Low Frustration</th>
<th>High Frustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation</td>
<td>Positive Politeness</td>
<td>Positive Politeness</td>
</tr>
<tr>
<td>Confidence</td>
<td>1) Bald, 2) Positive Politeness, 3) Negative Politeness</td>
<td>1) Positive Politeness, 2) Negative Politeness</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>Positive Politeness</td>
<td>1) Positive Politeness, 2) Negative Politeness</td>
</tr>
<tr>
<td>Performance</td>
<td>1) Bald, 2) Positive Politeness</td>
<td>Negative Politeness</td>
</tr>
</tbody>
</table>

**METHOD**

**Independent Variables**

The two independent variables were *Frustration* (low, high) and *Adaptation* (systematic, random). Frustration was manipulated by misaligning expectations of problem difficulty with reality, based on the theory that frustration comes from unfulfilled expectations (Berkowitz, 1989). Conversely, frustration may be diminished when people have lower level of expectation about their goal achievement (Bessiere et al., 2004). Recognizing a difference between the expected and actual level of difficulty can cause frustration (Hone, 2006; Glass, McGuinness, & Wolverton, 2008). All problems had a similar difficulty level, with historical Graduate Record Exam (GRE) correct answer rate of 30% – 40%. However, half of the problems were labeled as ‘easy’ even though they were just as difficult as problems labeled “hard”. In pilot tests, math problems of different difficulty levels and labels were given to participants, who self-reported their level of frustration on a 10-point scale. The final level of difficulty was chosen such that it produced enough frustration (when mislabeled as “easy”) to impact learning but not too high to cause the participant to give up.

Adaptation was manipulated by adjusting how the etiquette strategy for the next math problem was chosen: systematically or randomly. Systematic adaptation used the rule set described in the previous section. Random adaptation randomly chose etiquette strategies regardless of previous results. Systematic adaptation was compared to random adaptation rather than a no-feedback baseline since any feedback, no matter how poorly designed, could benefit
the learner due to the added information of the feedback as well as the possibility that changes in the feedback itself may cause the learner to engage with the material more. Previous work has established that feedback in the optimal etiquette strategy was significantly better than no-feedback (Yang & Dorneich, 2016). A random number generator set the transitions between etiquette strategies within the random trials. The strategy changed between each problem. It should be noted that in the random condition, there was a 25% chance that the etiquette strategy chosen for a math problem is the same one that would be chosen under the systematic condition. This may make differences harder to detect.

Hypothesis

- H1: Applying etiquette strategies based on the systematic rule set increases motivation, confidence, satisfaction, and performance more than randomly applying etiquette strategies.
- H2: Applying etiquette strategies based on the systematic rule set mitigates frustration more than random etiquette strategies.

Participants

A power analysis of the first 10 subjects indicated a minimum sample size of 21.75 to detect a significant effect for performance (power .90, alpha .05). Thirty-three university students (19 males, 14 females) participated in the experiment (demographics in Table 4), had normal or corrected-to-normal vision, and majored in STEM fields.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Mean</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>25.5</td>
<td>range: 19 – 31</td>
</tr>
<tr>
<td>Computer usage</td>
<td>8.0 hours/day</td>
<td>range: 3 – 16</td>
</tr>
<tr>
<td>algebra</td>
<td>8.2</td>
<td>SD = 1.5</td>
</tr>
<tr>
<td>geometry</td>
<td>7.0</td>
<td>SD = 1.9</td>
</tr>
<tr>
<td>trigonometry</td>
<td>6.8</td>
<td>SD = 1.8</td>
</tr>
<tr>
<td>calculus</td>
<td>7.8</td>
<td>SD = 1.8</td>
</tr>
<tr>
<td>statistics</td>
<td>6.6</td>
<td>SD = 2.5</td>
</tr>
<tr>
<td>probability</td>
<td>6.9</td>
<td>SD = 2.2</td>
</tr>
<tr>
<td>Last attended math class</td>
<td>2.6 years</td>
<td>Range: 1 – 5</td>
</tr>
</tbody>
</table>

Experimental Design

The experiment was a 2 (Frustration: low, high) x 2 (Adaptation: systematic, random), repeated measures, within-subject design, with one trial for each combination of independent
variables. Each trial contained five math problems, where the etiquette strategy varied either systematically or randomly between math problems, and the problems were all labeled either “hard” or “easy”. Trial order was counterbalanced using 4x4 Latin squares.

**Dependent Variables**

The dependent variables are described in Table 5.

**Table 5. The measurements for both independent variables verification and dependent variables.**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Metric</th>
<th>Measurement (Unit)</th>
<th>Frequency</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV Manipulation Check</td>
<td>TLX Subscale Frustration</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
<tr>
<td>Task Performance</td>
<td>Problem Solving Score</td>
<td>Score 0 – 10</td>
<td>After each math problem within trial</td>
<td>Objective</td>
</tr>
<tr>
<td>Motivation</td>
<td>Motivation Questionnaire</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
<tr>
<td>Confidence</td>
<td>Confidence Questionnaire</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>Satisfaction Questionnaire</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
<tr>
<td>Effectiveness when Targeting an</td>
<td>Change in level of learning factors</td>
<td>(-10) – (+10)</td>
<td>After each math problem within trial</td>
<td>Objective</td>
</tr>
<tr>
<td>Improvement in a Learning Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td>Effectiveness Questionnaire</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
<tr>
<td>Feedback Appropriateness</td>
<td>Appropriateness Questionnaire</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
<tr>
<td>Cognitive Workload</td>
<td>TLX Subscale Mental Demand</td>
<td>Scale 0 – 10</td>
<td>After each math problem within trial</td>
<td>Subjective</td>
</tr>
</tbody>
</table>

Independent Variable Manipulation Check. To confirm that the labeling of problems induced differing levels of frustration, participants rated their frustration on the NASA TLX (Hart & Staveland, 1988) Frustration subscale after each trial.
Motivation, Confidence, and Satisfaction. After each math problem, participants rated, on a 10-point scale:

- How motivated do you feel to continue working on tasks?
- How much satisfaction did you experience based on system’s feedback?
- How confident do you feel about your performance during the task?

Task Performance. The subject solved the math problem presented by the system using paper and pencil. Performance was graded (Table 6) immediately after each problem by the same experimenter who triggered feedback during the problem. The experimenter was the same for all subjects.

Table 6. The rubric to score problems.

<table>
<thead>
<tr>
<th>Score</th>
<th>Answer Sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Answer is correct with variables and equations demonstrated</td>
</tr>
<tr>
<td>7.5</td>
<td>Answer has correct equation but with calculation mistakes</td>
</tr>
<tr>
<td>5.0</td>
<td>Answer has correct solution approach such as setting up the variable from problems or drawing shapes based on given problems, but used wrong equations or no equations</td>
</tr>
<tr>
<td>2.5</td>
<td>If the participant tried to make variables or draw shapes but they were not the correct approach</td>
</tr>
<tr>
<td>0</td>
<td>If the answer sheet does not have anything correct</td>
</tr>
</tbody>
</table>

Effectiveness when Targeting an Improvement in a Learning Factor. The effectiveness and tradeoffs of the mitigation on targeted and non-targeted factors was measured by the difference in ratings between two consecutive problems for the four learning factors (ΔMotivation, ΔConfidence, ΔSatisfaction, ΔPerformance).

Feedback Appropriateness, Effectiveness, Cognitive Workload. Cognitive workload was measured with the mental demand subscale of the NASA TLX. Two other constructs were rated on a 10-point scale:

- Was feedback from the system appropriate?
- Was feedback from system effective to solve the task?
Procedure

This research complied with the APA Code of Ethics and was approved by the Institutional Review Board at Iowa State University (16-004). After informed consent, participants were provided training and practice to solve math problems until comfortable.

During trials, the experimenter observed the participant and triggered feedback based on the rubric (Table 2). This ensured all feedback was triggered correctly, and effects uncovered were solely due to etiquette strategies and not confounded with triggering the wrong feedback. After every math problem, participants rated motivation, confidence, and satisfaction. The experimenter calculated their performance score. The experimenter entered the ratings and performance score into a rule-set calculator (excel file) to determine the etiquette strategy for the next math problem. This procedure took less than one minute. The etiquette strategy only changed between math problems. After every trial (5 math problems) participants rated workload, feedback effectiveness, and feedback appropriateness. A post-experiment survey gathered formative feedback. The debriefing revealed that the true goal was to study etiquette strategies, not mathematics ability.

Data Analysis

The Shapiro-Wilk test found that normality assumptions were not met for most metrics, so non-parametric Wilcoxon tests were used. To answer the question, “Was the targeted learning factor improved by the system?” a Wilcoxon analysis determined the effect of adaptation and frustration on the targeted learning factor. To answer the question, “What also happened to the non-targeted learning factors?” a nonparametric Steel-Dwass test was conducted on each pairwise combination of learning factors for each adaptation type. Wilcoxon results are reported as highly significant for a significance level alpha <.001, significant for alpha <.05, and marginally significant for alpha <.10 (Gelman, 2013). In addition, since Wilcoxon tests cannot calculate an interaction effect between two factors, the Aligned Rank Transform (ART) method was used to analyze interaction effects in nonparametric factorial data (Wobbrock, Findlater, Gergle, & Higgins, 2011). ART aligned and ranked data before using regular F-tests. Cohen’s d measured effect size of the mean difference between two groups in standard deviation units, and were reported as small (.20 < d <.50), medium (.50 < d <.80), and large (d >.80). A Pearson product-moment correlation coefficient \( r_s \) was calculated to understand the association between
demographic data (age, skill level, years since last math class) on the learning factors. None of the demographic factors has any correlation (\(|r_s| > 0.3\)) to motivation, confidence, satisfaction, or performance scores.

**RESULTS**

*Independent Variable Manipulation Check*

Frustration rating was significantly \((Z = 13.6, p < .001, d = 1.09)\) higher for the high frustration (“easy” label) condition than the low frustration (“hard” label) condition (Figure 3). The main effect of adaptation was not significant, \(Z = 0.41, p = .68\). This verifies that the manipulation of frustration was effective.

![Figure 3. Mean and standard error of TLX frustration rating.](image)

*Effectiveness when Targeting an Improvement in Learning Factor*

Given that any learning factor could be targeted for a math problem, the number of data points for each factor may be different.

*Targeting Motivation.* Figure 4 illustrates the change in motivation (\(\Delta Motivation\)) when targeting motivation. The main effect of adaptation was significant, \(Z = 2.14, p = .033, d = 0.50\). Systematic adaptation resulted in a larger \(\Delta Motivation\) \((M = 2.8, SD = 4.2, N = 77)\) than random adaptation \((M = 1.0, SD = 3.1, N = 62)\). The main effect of frustration on \(\Delta Motivation\) was not significant, \(Z = 0.30, p = .76\). The interaction effect was non-significant, \(F(1,16) = 0.66, p = .43\).
Figure 4. Mean and standard error of motivation rating change.

Figure 5 illustrates how learning factors were affected when motivation was targeted. Since the main effect of frustration was not significant, the data were collapsed across this variable. For the systematic adaptation, Figure 5 indicates significant pairwise differences between groups when they do not share a capitol letter. ∆Motivation was significantly larger than the other three learning factors, with a large average effect size of $d = 0.97$. For the random adaptation, Figure 5 indicates significant pairwise differences between groups when they do not share a lower-case letter. ∆Motivation was significantly larger than ∆Satisfaction and ∆Performance, with a small average effect size of $d = 0.45$.

Figure 5. Mean and standard error of ∆Motivation with systematic (n=77) and random adaptation (n=62) for both the targeted (motivation) and non-targeted learning factors. A significant pairwise difference between groups is indicated when they do not share a capitol letter (Systematic) or lower case letter (Random).
Targeting Confidence. The main effect of adaptation was significant, $Z = 2.72, p = .006, d = 0.56$ (Figure 6). Systematic adaptation resulted in a larger $\Delta$Confidence ($M = 4.3$, $SD = 3.52$, $N = 60$) than random adaptation ($M = 2.0$, $SD = 4.5$, $N = 58$). The main effect of frustration on $\Delta$Confidence was not significant, $Z = 0.78, p = .44$. The interaction effect was non-significant, $F(1,19) = 0.41, p = .53$.

![Figure 6. Mean and standard error of confidence rating change.](image)

For systematic adaptation, $\Delta$Confidence was significantly larger than the three non-targeted learning factors, with a large average effect size of $d = 1.09$ (Figure 7). For random adaptation, $\Delta$Confidence was significantly larger than three non-targeted learning factors, with a moderate average effect size of $d = 0.58$.

![Figure 7. Mean and standard error of motivation with systematic (n=60) and random adaptation (n=58) for both the targeted (confidence) and non-targeted learning factors. A significant pairwise difference between groups is indicated when they do not share a capitol letter (Systematic) or lower case letter (Random).](image)
**Targeting Satisfaction.** The main effect of adaptation was marginally significant, $Z = 1.89$, $p = .058$, $d = 0.42$ (Figure 8). Systematic adaptation resulted in a larger $\Delta$Satisfaction ($M = 4.0$, $SD = 3.9$, $N = 60$) than random adaptation ($M = 2.3$, $SD = 4.1$, $N = 59$). The main effect of frustration on $\Delta$Satisfaction was not significant, $Z = 0.07$, $p = .94$. The interaction effect was non-significant, $F(1,10) = 0.11$, $p = .75$.

![Figure 8. Mean and standard error of satisfaction rating change.](image)

For systematic adaptation, $\Delta$Satisfaction was significantly larger than the three non-targeted learning factors, with a large average effect size of $d = 1.03$ (Figure 9). For random adaptation, $\Delta$Satisfaction was significantly larger than three non-targeted learning factors, with a moderate average effect size of $d = 0.58$.

![Figure 9. Mean and standard error of motivation with systematic (n=60) and random adaptation (n=59) for both the targeted (Satisfaction) and non-targeted learning factors. A significant pairwise difference between groups is indicated when they do not share a capital letter (Systematic) or lower case letter (Random).](image)

**Targeting Performance.** The main effect of adaptation was significant, $Z = 2.50$, $p = .012$, $d = 0.47$ (Figure 10). Systematic adaptation resulted in a larger $\Delta$Performance ($M = 2.5$, $SD = 3.6$, $N = 62$) than random adaptation ($M = 0.73$, $SD = 3.6$, $N = 82$). The main effect of frustration on
ΔPerformance was not significant, $Z = 0.32, p = .75$. The interaction effect was non-significant, $F(1,24) = 2.76, p = .11$.

![Figure 10. Mean and standard error of performance rating change.](image)

For systematic adaptation, ΔPerformance was significantly larger than the three non-targeted learning factors, with a large average effect size of $d = 0.87$ (Figure 11). For random adaptation, ΔPerformance was significantly larger than ΔConfidence and ΔSatisfaction, with a small average effect size of $d = 0.41$.

![Figure 11. Mean and standard error of motivation with systematic (n=62) and random adaptation (n=82) for both the targeted (Performance) and non-targeted learning factors. A significant pairwise difference between groups is indicated when they do not share a capital letter (Systematic) or lower case letter (Random).](image)

**Feedback Effectiveness**

The main effect of adaptation on feedback effectiveness was not significant, $Z = 0.88, p = .38$ (Figure 12). The main effect of frustration on feedback effectiveness was marginally significant, $Z = 1.80, p = .072, d = .21$. The effectiveness rating was higher in the high frustration condition ($M = 5.8, SD = 3.7$) than in the low frustration condition ($M = 5.3, SD = 3.7$). The
interaction effect was non-significant, $F(1,29) = 0.15, p = .70$. There was no significant effect of targeted learning factor on feedback effectiveness, $Z = 3.53, p = .31$.

![Figure 12. Mean and standard error of feedback effectiveness.](image)

**Feedback Appropriateness**

The main effect of adaptation on feedback appropriateness was not significant, $Z = 0.88, p = .38$ (Figure 13). The main effect of frustration on feedback appropriateness was marginally significant, $Z = 1.80, p = .072, d = .19$. The appropriateness rating was higher in the high frustration condition ($M = 6.3, SD = 3.5$) than in the low frustration condition ($M = 5.8, SD = 3.6$). The interaction effect was non-significant, $F(1,30) = 1.81, p = .19$. There was no significant effect of targeted learning factor on feedback appropriateness, $Z = 1.32, p = .72$.

![Figure 13. Mean and standard error of feedback appropriateness.](image)
Cognitive Workload

The main effect of adaptation on the TLX mental demand rating was not significant, \( Z = 0.78, p = .44 \) (Figure 14). The main effect of frustration on mental demand was significant, \( Z = 3.36, p < .001, d = 0.49 \). Mental demand was higher in the high frustration condition (\( M = 3.2, SD = 2.2 \)) than in the low frustration condition (\( M = 2.4, SD = 4.5 \)). The interaction effect was non-significant, \( F(1,33) = 1.01, p = .32 \).

![Figure 14. Mean and standard error of TLX mental demand.](image)

DISCUSSION

Systematically adapting interaction style based on etiquette strategies significantly influenced motivation, confidence, satisfaction, and performance. Hypothesis H1 was fully supported. The targeted learning factor under systematic adaptation increased significantly more than under random adaptation. Hypothesis H2 was partially supported. Systematic adaptation significantly improved targeted learning factors in both high and low frustration conditions, with the same level of improvement regardless of frustration level. However, frustration did not affect the random condition significantly either. The aligns with the observation of Robison et al (2009) that negative emotions such as frustration may benefit from interventions even when inappropriately delivered. However, this could also be because the high frustration condition averaged a rating of four out of 10, possibly a level not high enough to significantly affect the learning process. Although even at this moderate level of frustration, there was a marginally significant, albeit small effect on the appropriateness and effectiveness ratings of the feedback.

Frustration may affect performance (Libb, 1972) because it indicates dissatisfaction related with encountered difficulties in learning (Radford, 2015). Frustration induced in this work was one of a mismatch of expectations to actual difficulties. Future work will study the effect of
higher levels of frustration on the learning factors. Future work can also explore other sources of frustration including task difficulty itself (Rosenzweig, 1938), time pressure (Wahlström, Hagberg, Johnson, Svensson, & Rempel, 2002), ineffective or irrelevant feedback (McKinney, 1933; Baylor & Rosenberg-Kima, 2006), or human-computer interaction issues (Klein et al., 2002; Powers et al., 2011).

The study also investigated potential adaptation tradeoffs. Systematic adaptation improved targeted leaning factors significantly more than non-targeted learning factors. An ITS system would have to decide when a small negative change in the non-targeted factor was acceptable for the larger gains in the targeted factor. Previous work had established that “polite” feedback was more effective than “direct” feedback for some measures of motivation and performance (Wang et al., 2005; McLaren et al., 2011). This work demonstrates that “polite” feedback can systematically be adapted during instruction to individually target and improve motivation, confidence, satisfaction, and performance, where systemic application outperforms random (but still polite) application.

However, the random adaptation also significantly improved the targeted learning factor compared to non-targeted factors. But where effect sizes for systematic adaptation were large, effect sizes in the random adaptation were small. There are several possibilities why there was an improvement in random adaptation at all. Post hoc analyses determined that 21.2% of the time the random choice of etiquette strategy matched the result that would have been obtained by using the systematic rules. A second possibility is that the activation caused by any change at all resulted in improvement. However, none of the non-targeted learning factors showed improvements and so this possibility is less likely. Finally, there might be an adjustment towards the mean, where the learning factor most in need of improvement (the targeted factor) would improve the most in the next trial, no matter what etiquette strategy was applied. If this was the case then the random and systematic conditions would see similar levels of improvement in the targeted factor. However, the magnitude of the improvement under systematic adaptation was 2.5x larger (range: 1.7 – 3.4) compared to random adaptation.

A good human tutor is aware of learner emotional state and adapts their interaction style to support factors that underlie performance such as a motivation, confidence, or satisfaction. However, it can be difficult for a human tutor to determine how best to respond to a student’s emotional needs, depending on the human tutor’s ability to both detect the emotion correctly, and
then have a strategy on how to respond (Petrovica, 2013). There is a lack of a theoretical framework to determine how an ITS can use emotions in learning situations (Petrovica, 2013). Developing affect-aware ITS are still in early stages, and much more work is needed to understand the full impact of the use of emotions in tutoring systems (D’Mello & Calvo, 2012). Leveraging previous work in human-human and human-computer etiquette, this work aims to provide mechanisms to enable an ITS to adapt its interaction style to impact the learning factors that support performance as identified by Keller (2009). The evaluation results demonstrate that not only do changing etiquette strategies have a differential effect on student learning factors, but that those differential effects can be used systemically to target and improve learning factors. However, many research questions remain. Changes in etiquette strategy happened between each problem, but human tutors may be able to change their strategies during a problem. Thus, future work may look at moment-to-moment detection of human affect to trigger changes in real-time. Additionally, this work focused on testing the effectiveness of the adaptions and assumed “perfect triggering” to initiate feedback by having the experimenter decide when to trigger feedback instead of sensors to detect learner state. This ensured that any effects could be attributed to the etiquette strategies. However, a fully closed-loop affect-aware ITS would automatically detect frustration, learning factors, and problem-solving difficulties in real time. Thus, two lines of research are needed to address this gap: 1) automated detection of frustration and learning factors and 2) ITS logic that incorporates imperfect detection. Using sensors to detect emotional state is an active area of research, and includes video of facial expressions (Loijens et al., 2012), facial electromyography (Hazlett, 2003), or skin conductance (Boucsein, 2012). Every detection system has some level of uncertainty (i.e. “imperfect triggering”), thus further work is needed to understand how the trustworthiness, uncertainty, and fuzziness of emotion-detection systems impact overall ITS effectiveness (Landowska, 2013). For instance, there may be an interaction between inappropriate feedback (due to an incorrect detection of emotional state) and motivation and trust. An ITS that attempts to display higher levels of emotional intelligence and support may invoke higher expectations on the part of learners. If the systems “get it wrong” then trust may erode, which may lead to lower motivation, confidence and satisfaction. Understanding what the threshold of accuracy is needed before providing the wrong interaction style is worse than making no change at all is an open question that future research must address.
Although this study looked only at modifications of interaction (how), fully formed adaptive ITSs would still consider modification of content (what) and modification of task scheduling (when). For instance, confidence could be targeted by presenting easy questions in a lesson first, before progressing to more difficult items. Future work will study the interplay of the timing, content, and style of adaptive feedback to improve learners cognitive and affective states. Individual differences of the participants must be considered. The ultimate goal is to develop a library of possible options (of feedback content and etiquette strategy) and a rule set that triggers the appropriate feedback to best meet the learning objectives. Finally, this experiment used only math problems. The type of task may influence which etiquette strategy is optimal under what conditions. A future area of research would be to expand this adaptive interaction style approach to non-STEM fields.

**Acknowledgements**

The authors would like to thank Mariangely Iglesias-Pena, David Montealegre, Jordan Zonner, and Maria Dropps for supporting experiment design and data analysis. The authors would like to thank Leslie Potter and Dr. Caroline Hayes for feedback on the manuscript. This material is based in part upon work supported by the National Science Foundation under Grant No. 1461160.

**References**


Adaptive Etiquette Strategies in ITSs


Koedinger, K., & Tanner, M. (2013). Things you should know about intelligent tutoring systems. EDUCAUSE Learning Initiative (ELI).


Adaptive Etiquette Strategies in ITSs


Biographies

Dr. Euijung Yang is a human factors engineer and user experience researcher at ConnectFive. She earned her PhD in human factors in the Industrial and Manufacturing Systems Engineering Department at Iowa State University in 2016. She earned her BS and MS at in Industrial Engineering at Soongsil University.

Dr. Michael C. Dorneich is an Associate Professor at Iowa State University in the Industrial and Manufacturing Systems Engineering Department since 2012. Previously he was a Principal Research Scientist in the Human Centered Systems department at Honeywell Laboratories. He earned his PhD in Industrial Engineering in the human factors program at the University of Illinois at Urbana-Champaign in 1999.