RICE UNIVERSITY

The impact of Feedback Tone, Grammatical Person and Presentation Mode on Performance and Preference in a Computer-based Learning Task.

by

Sebastian Thomas

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

David Lane, Associate Professor,
Psychology, Statistics and Management,
Chair

Michael Byrne, Professor, Psychology

Phillip Kortum, Assistant Professor,
Psychology

H. Albert Napier, Professor of
Management

HOUSTON, TEXAS
January 2013

RICE UNIVERSITY

The impact of Feedback Tone, Grammatical Person and Presentation Mode on Performance and Preference in a Computer-based Learning Task.

by

Sebastian Thomas

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

David Lane, Associate Professor,
Psychology, Statistics and Management,
Chair

Michael Byrne, Professor, Psychology

Phillip Kortum, Assistant Professor,
Psychology

H. Albert Napier, Professor of
Management

HOUSTON, TEXAS
January 2013

ABSTRACT

The impact of Feedback Tone, Grammatical Person and Presentation Mode on Performance and Preference in a Computer-based Learning Task.

by

Sebastian Thomas

Politeness is a part of student-tutor interactions and research in affective computing has shown that this social convention may also be applicable when a computer plays the role of tutor. This study sought to build on previous work that examined the effect of the politeness of computer feedback through the application of social and cognitive theories. Employing a mixed-factor design, a sample of 150 college students completed a multiple cue probability learning task (MCPL) on a computer that provided feedback phrased in one of three different tonal styles (joint-goal, student-goal and bald-on-record). Feedback tone was a within-subjects factor. Subjects received feedback as either text or as audio. Audio feedback was a between-subjects factor and was delivered in one of four different modes male/female human voice or a male/female synthesized voice. The study found gender differences in tone preference as well as a possible impact of the Tone x Mode interaction on learning. Specifically, men were more likely than women to prefer the student-goal style feedback prompts. It is hoped that this research can provide additional insight to designers of learning applications when they are designing the feedback mechanisms that these systems should employ.

Acknowledgments

I would firstly like to thank my advisor David Lane, whose guidance and feedback during all phases of this research was instrumental in its completion. I also appreciate the insight provided by my other committee members: Michael Byrne, Phillip Kortum and Al Napier. This input greatly enhanced the quality of this research. I would also like to acknowledge the voice actors and raters that helped with the selection and production of voice prompts. A final “thank you” goes to my research participants who were kind enough to participate in this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Feedback plays an essential role in successful learning environments. To illustrate, imagine a classroom in which a teacher provides instruction, administers tests and asks students questions, but never pauses to answer questions posed by students, grade tests or discuss the accuracy and merits of the answers given by students. Although certainly not impossible, it is not difficult to envision that learning would be very difficult in such a circumstance.

Feedback can be defined in many ways and can go in multiple directions. In the above example for instance, although the teacher may not be giving overt feedback to the students, they are giving the teacher feedback with their answers to the teacher's questions. This feedback in turn has the potential to facilitate learning by allowing the teacher to modify his/her method of instruction. The student-tutor interaction is governed by a wide set of social conventions and mores all of which have the potential to influence learning and the general motivation of the learner. This study sought to examine the

extent to which these conventions may apply in a circumstance where the human tutor is replaced by a computer.

1.1. Formative Feedback

Before discussing how technology can be used to provide feedback, a formal definition of how feedback was understood in this study is provided. This research focussed only on feedback in the teacher-to-student direction or what has been referred to in the literature as formative feedback. Formative feedback as defined by Shute (2008) is "…information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning." There are multiple factors that contribute to the effectiveness of formative feedback and there have been multiple reviews of the literature and meta-analyses that aggregate these effects in an attempt to build a framework to explain them (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Kulhavy & Stock, 1989; Shute 2008). Hattie and Timpereley (2007) in particular, proposed a model of feedback in which to be effective, feedback must help to answer three questions for the learner: (1) What are my goals?, (2) What progress am I making towards them?, and (3) What do I need to do to make better progress? The model goes on to show that feedback can operate at four different levels and the effectiveness of answers to the three questions depends on the level at which feedback occurs. The four levels are categorized based on their task, procedural, cognitive and personal properties. These levels are not mutually exclusive as feedback given to the learner often possesses properties at two or more of these levels. The four levels are described in some detail.

Task level feedback, which has also been referred to as corrective feedback, provides information on how well a task is understood and performed and it is probably the most common form of feedback that a teacher gives to a student. Feedback about whether an answer is correct or incorrect is typical at this level. Hattie & Timperley (2007) point out that task level feedback may also include instructions on how to acquire additional information that can be used to improve an answer. Evidence that task feedback is an effective teaching tool is substantial and has existed for a long time. Lysakowski and Walberg (1982), in a compilation of 20 studies utilizing corrective feedback the earliest of which dated back to 1967, report a mean effect size of 0.94 with a standard deviation of 1.91. Hattie and Timperley (2007), in a more up-to-date review of the literature report somewhat smaller, but still moderate effects ranging from 0.25 to 0.43.

Task level feedback's effectiveness is moderated by several factors, one of which is the amount of experience that learners have with the material that they are trying to learn. Moreno (2004) found that corrective feedback was less effective than more procedural style explanatory feedback in terms of transfer, motivation and perceived helpfulness for students who were novices in a discovery learning environment.

Kantner and Lindsay (2010) in a series of experiments based on variations of an old/new recognition task found that corrective feedback significantly improved response bias (the ability to know which response is most likely) particularly when the base rate of old items differed greatly to that of new items. However, not only did these experiments fail to find improvement in the recognition sensitivity (the ability to recognize previously seen words), one experiment actually yielded a significantly negative impact of feedback

on sensitivity. Kanter and Lindsay (2010) explain their findings by suggesting that the "inferential decision-making processes" utilized by the "recognition system" are mostly inaccessible to the potential positive effects of corrective feedback. Research has also shown that corrective feedback's impact can change depending on when it is introduced in the learning process. In one such study Corbalan, Paas and Cuypers (2010) found that students solving linear algebra problems on a computer had better learning outcomes and were more motivated when they received immediate feedback after each solution step in solving an algebra problem (step-wise) compared to receiving feedback only after submitting a final solution.

Process level feedback provides information regarding the underlying rules and relationships upon which the, to be learned, material is founded. To contrast between the task and process levels, consider a student receiving feedback about his/her incorrect answer to an arithmetic problem. At the task level, feedback may be along the lines of "your answer is two points too high" while feedback targeting the process level could be phrased as "when solving these types of problems, you should always solve what is in the parentheses first." Process level feedback is more commonly referred to as cognitive feedback in the literature and research has shown that it can often be more effective than task level feedback. It should be noted that process and task level feedback can be effectively presented together in most paradigms.

Feedback pertaining to process is particularly helpful for novice users. As discussed earlier for instance, Moreno (2004) found that domain novices benefited more from process feedback than from task feedback. On the other hand, process feedback can be less effective for domain experts. As evidence of this, Montanzemi and Gupta (1998)

tested the impact of cognitive feedback via a decision support system for experienced operators in an industrial machinery context and found that feedback had no effect on task performance.

The third level of feedback is achieved through self-regulation and can be thought of as a type of metacognition where the student reflects on his/her performance of a task and in so doing is able to detect errors and adjust subsequent performance. Feedback at this level is internal to the learner and as such fell somewhat out of the scope of this study since it does not flow in the tutor to learner direction. It has been shown however, that self-regulation can affect the effectiveness of feedback at the other levels and thus merits some discussion. In a review of research involving feedback and self-regulation, Butler and Winne (1995) make the argument that self-regulation is an inherent aspect of learning and that the effective learner also has to be an effective self-regulator. In support of the idea that the interplay of feedback at the four levels can influence their effectiveness, these researchers also posit that while self-regulation can lead to issues when the learner has formed goals and strategies that are contrary to the actual learning objectives, the learner can reduce these inaccuracies by receiving task and process level feedback.

Task and process level feedback are often accompanied by feedback directed towards the learner as a person, typically in the form of praise or criticism. Hattie and Timperley (2007) in their review, argue that this type of feedback is often the preferred form of feedback used by teachers although it is generally ineffective because it typically provides no, or only marginally useful task relevant information. Hawley et al. (1984) explain that praise, while generally intended as a reinforcer for desired student behavior,

is often more of a response by the teacher that is elicited and reinforced by the students. A statement such as “good job you are doing really well” gives the student very little information about which specific aspects of the task are being done well and whether there is any room for improvement. Nass and Yen (2010) take a similar view regarding praise in that they report having seldom seen positive outcomes related to feedback in the form of praise in the research conducted by themselves and their associates. They argue that individuals do not “deeply” think about received praise and so have difficulty remembering the specifics of it. Conversely, criticism, according to Nass and Yen (2010) is processed more deeply and may therefore lead to better recall of feedback. The researchers do mention a caveat where criticism, because it uses more cognitive resources, may affect the recipient’s recall of information received preceding criticism (retroactive interference).

1.2. Politeness Theory

There are certain social conventions that affect interactions between a tutor and student and these can have an effect on how feedback is delivered and accepted. Of particular relevance to this research is politeness and how it can be used to affect the way in which feedback is received by the learner. Politeness represents a significant aspect of how human beings interact with each other. Brown and Levinson (1987) developed a cross-cultural theory that sought to describe how politeness is used in social interactions. They posited that individuals try to maintain “face” when they interact with each other. “Face” is from the folk term “losing face” and the theory holds that individuals try to manage two types of faces in social interactions: positive face and negative face. Positive

face refers to the desire of individuals to be viewed favorably by those that they view as being important while negative face refers to the desire of individuals to not be impeded by others.

Within the paradigm of the learning task that was used for this research (to be discussed later), a feedback prompt such as "you are over weighting cue 1, you must focus more attention on the other cues" would be viewed as impolite because it entails two different face threats. First, there is a threat to positive face at the beginning of the prompt which criticizes the user's performance and second, there is a threat to negative face in the second half of the prompt which prescribes a specific action that the user "must" take to rectify their mistake. Politeness theory holds that the greater the number of face threats in a statement the more impolite it is viewed as being.

One of the tenets of Brown and Levinson's theory is that speakers manage face threats by employing politeness strategies. The authors classify these strategies into four broad categories: off-record, bald-on-record, negative politeness and positive politeness. An off-record statement is one in which the speaker's intentions are ambiguous leaving room for interpretation and often conveying meaning that goes beyond what is explicitly stated. An example of an off-record statement within the context of feedback from a computer application could be: "The system indicates that the entered value is incorrect." This statement introduces ambiguity by referring to the "system" and not directly to the user. It may also be clear from this simple example that the intent is for the user to correct the value entered, but this intent must be inferred as it is not stated directly. Levinson and Brown (1989) do point out that there are instances where ambiguous statements can cease to be off-record given social conventions that may be present in particular cultures. A

statement such as “can you turn the music down” is phrased as a question but would be understood to be a request by English speakers. Politeness theory sees these types of utterances as utilizing “conventional indirectness.”

Conversely, a bald-on-record statement is one in which the speaker clearly and concisely communicates intent. These types of statements are generally seen as less polite than off-record statements because no attempt is made to protect the addressee’s “face.” A bald-on-record rewording of the previous example could be: “The value you entered is wrong, you need to correct it.” This statement goes “on-record” because its intent is stated explicitly and is blunt (bald) in its phrasing.

A speaker may also choose to go “on-record” but maintain some regard for the addressee’s “face”. Levinson and Brown (1989) categorize these strategies into positive and negative politeness, based on which type of threat a strategy is intended to mitigate. A speaker may choose to employ multiple strategies in a single statement and these strategies may address both types of face threats. Continuing with the previous example, an on-record and more polite rephrasing of the statement could be: “It looks like the value you entered was not quite right; you should try a different value.” This prompt employs both positive (first half) and negative (second half) politeness in an effort to minimize threats to face. More specifically, the insertion of the modifiers “it looks like” and “not quite” as well as the modal verb “should” change the statement from a criticism and command (as was the case in the bald-on-record form) into more of an observation and recommendation.

Politeness theory points to a variety of factors that determine the extent to which individuals use the various politeness strategies when interacting with each other. Brown and Levinson divide these factors into "intrinsic payoffs" and "relevant circumstances." Payoffs are the perceived benefits that the speaker derives from a given strategy. An off-record strategy for instance, may allow the speaker to appear more tactful and reduce accountability for statements made. Both of these payoffs are useful when the speaker is delivering information that can be perceived as being critical of the recipient. On the other hand, payoffs for a bald-on-record strategy might be greater efficiency as well as exerting greater pressure on the addressee to act.

Relevant circumstances are another factor discussed by Brown and Levinson and refer to a set of social variables whose values are determined by the speaker and the addressee. Specifically, these factors include the perceived power relationship between the speaker and the addressee as well as the social distance between them. In a tutor-student relationship the power typically resides with the tutor and this will affect how politeness strategies are utilized. In general, a speaker who is in a more powerful position has more freedom in the choice of communication strategy.

Subsequent research has adapted Brown and Levinson's original theory, the most commonly used of which divide politeness strategies into eight categories (e.g. Mayer et al., 2006; Wang et al.,2008). The eight strategies are shown in Table 1 below along with an example of each.

Strategy	Example
Bald-on-record	You must correct the entered value.
Conventional-indirectness	The system indicates that the entered value is incorrect.
Request	I would like you to correct the value you entered.
Question	Could you correct the value you entered?
Tutor -goal	I would correct the entered value.
Student-goal	You may want to correct the value you entered.
Joint-goal	We should correct the value we entered.
Socratic-hint	Do you want to correct the value you entered?

Table 1 Politeness strategies and examples

The first two strategies listed in Table 1 can be linked directly to Brown and Levinson's original theory and have already been discussed. The six remaining strategies can be employed with either positive or negative politeness. A *request* is phrased in such a way as to convey the speaker's desire while a question is an indirect query that hints at an action. The three goal strategies (tutor, student and joint) should be fairly self-explanatory based on the examples, in that statements are phrased in such a way that makes the required action a goal of the tutor, student or both. The socratic hint is a more subtle goal strategy where the student's goal is presented as a suggestion in the form of a question.

1.3. Affective computing and formative feedback

It is not difficult to imagine that formative feedback would be as important in a circumstance in which teaching is conducted entirely via a computer application as it would in any other learning environment. What is more difficult to gauge is whether the quality of learning with a piece of software can be affected by the degree to which the social conventions that govern more traditional teacher-learner interactions are maintained. The increasing interest in, and use of technology in society has led to numerous avenues of research some of which have direct implications for learning. One such area is affective computing which examines the role of computers as social actors and the extent to which the social mores that govern human-human social interaction can be applied to human-computer interactions. Research in this area has repeatedly demonstrated that individuals in a variety of contexts in which they interact with computers, react in ways that one would only expect if the computer's role were being performed by a human actor. Examples include being polite and trying not to be overly critical when evaluating the performance of a computer used to complete a task (see Nass & Yen, 2010). Affective computing researchers are careful to point out that while individuals may to some degree apply social rules in human-machine interactions, these individuals, when asked, do not consciously ascribe any human properties to these computers (see Nass & Moon, 2000).

Given evidence that computers may have a role as social actors, politeness theory may, by extension, have implications for the design of teaching applications. Research in affective computing has shown that users' impressions of the computers they interact with can be changed by the type of language used in the prompts given by the computer.

For example, Nass *et al.* (1995) compared a computer system that used dominant-language prompts to one that used submissive-language prompts and found that subjects were more likely to prefer the computer whose personality was closer to their own. In this study, Nass *et al.* (1995) categorized personality as being either dominant or submissive based on a subscale from the Bem sex-role inventory (Bem, 1974). This scale has been found to correlate with other measures of dominance. Subjects scoring higher on dominance preferred direct feedback whereas those with lower scores preferred more submissive prompts.

Mayer *et al.* (2006) used this theory to develop "polite statements" that could be used for computer-based tutors to increase the social sensitivity of educational software. They found that subjects were able to discriminate between different levels of politeness among a set of computer application style prompts. Moore *et al.* (2004) applied principles from Brown and Levinson's theory to develop a model that could be used to generate tutorial feedback for a basic electronics tutorial that was comparable to the feedback provided by human tutors. As further evidence of the importance of social cues when providing feedback, Klein, Moon and Picard (2002) found that subjects used a frustrating computer longer when they were able to interact with an electronic agent that provided "active emotion support" via onscreen text.

The relationship between learning outcomes and feedback tone has not been studied extensively. Wang *et al.* (2008) found better learning when subjects received feedback that was polite rather than direct. These researchers used a Wizard-of-Oz paradigm in which feedback in one condition was polite and sought to reduce face threat whereas in the other condition feedback was direct with minimal politeness. Subjects

interacted with an online factory modeling and simulation application and were required to use it to forecast demand, develop a production plan and a process schedule for a virtual factory. In addition to better overall performance for subjects that received polite feedback, this study also found that subjects who reported that they preferred indirect help performed better in the polite feedback condition than the direct condition. This difference was not observed among subjects that reported a preference for direct feedback.

Mclaren, Deleeuw and Mayer (2011) also found a politeness effect on learning with a computer-based tutor. In this study, participants with varying degrees of chemistry knowledge learned about and solved stoichiometry problems and received direct or polite feedback as they solved these problems. The researchers found a learning interaction between politeness and the level of chemistry knowledge of participants, where those with low knowledge had better learning outcomes using an interface with polite feedback while high knowledge participants had similar learning outcomes regardless of the tone of the interface used. These researchers explain their findings in terms of the social agency principle which states that a learner will make a greater effort to understand a tutor if he/she is viewed as a social partner. Such a view is more likely to occur if the tutor employs a more polite conversational style particularly with low knowledge learners.

1.4. Grammatical Person and Feedback

One of the nuances of language that can be employed by teachers in giving feedback to their students is grammatical person. This is apparent in politeness theory

where changing the grammatical person of a statement can influence its perceived politeness. More specifically feedback statements in the second-person singular (e.g. "you did something wrong") are typically thought to be less polite than statements in the first-person plural (e.g. "we did something wrong"). Although it is covered indirectly in research involving politeness and learning, there is very little research specific to the impact of grammatical person *per se*. Brinko (1993) does mention grammatical person regarding teaching in general, by stating her observation from years of teaching experience that positive feedback is more effective in the second-person while negative feedback is more effective in the first or third-person. Her observation, which is in line with politeness theory, is that the grammatical second-person personalizes positive feedback and makes it more powerful. Conversely the grammatical first or third-person depersonalizes negative feedback which in her view makes it less accusatory and thus more palatable.

There is some anecdotal evidence that grammatical person may be important to the learning process especially with regard to computer feedback. Consider for instance a computer-generated feedback prompt that uses the first-person-plural. In a previous study Thomas and Lane (2010), found that when asked for their thoughts on the feedback prompts they received from a computer during a learning task, subjects often commented that the use of the pronoun "we" by the computer was "strange/weird". Shneiderman et al. (2009) cautions against the use of first-person pronouns in computer interfaces as, in his view, this type of wording can be distracting as well as confusing for users. One of the major goals of the research presented here was to provide more empirical evidence regarding the impact of grammatical person in the computer-generated feedback prompts.

1.5. Auditory feedback

Up to this point, the focus of this review has been on text-based feedback delivered to the learner by a computer via on-screen prompts and the role of social convention in this process. It is important to consider how these conventions apply when using auditory feedback. Generally speaking, memory researchers have uncovered fairly stable modality effects associated with the presentation of text as audio where recall is typically higher than text presented visually. Furthermore, text presented concurrently in both modalities has better recall than text presented in only one. Penney (1989) reviews early studies that demonstrate these effects. The dual-processing theory of working memory (Baddeley, 1992; Pavio, 1986) has been used to explain these findings. This theory describes working memory as having two separate stores: the phonological loop - which stores speech, and the visuospatial sketch pad -which is a visual image store. Given these separate stores it is often beneficial for information to be presented in multiple mediums to take advantage of both types of processing.

Modality effects have also been demonstrated with multimedia learning. One of the first published studies on this effect, Moreno and Mayer (1999), found that participants who were presented with narrated text in a meteorology themed learning application had better verbal recall, transfer and visual-verbal matching when text was narrated rather than presented visually. It should be noted that these effects have not always been replicated outside of laboratory settings. Tabbers, Martens and van Merriënboer (2004) administered a multimedia lesson on instructional design to students as a part of their course on the same subject and found that replacing visual text with narrated text actually led to lower retention and transfer. These researchers hypothesized

that factors such as pacing may have contributed to this result. The visual text interface was system paced while subjects could navigate at their own pace in the audio feedback interface. The researchers also pointed out that it took some time for the audio clips to download and that this may have impacted results as well. Ginns (2005) conducted a meta-analysis of studies examining the modality effect in multimedia learning and found that it occurred fairly consistently across 43 studies. Ginns (2005) also found that pacing moderated the size of the modality effect.

1.5.1. Male versus Female voices

The choice of voice gender for computer systems has been a fairly common subject of discussion and study. Anecdotally, one could argue that female voices have been more commonly used in the higher profile computer systems, with Apple's Siri being a good example. A Time article conjectured that one reason for the predominance of female computer voices could lie with Hollywood where male voices have been used to portray menacing computer systems such as HAL in "2001: A Space Odyssey." This contrasts with female voices which tend to be used to portray more supportive systems such as the onboard computer from "Star Trek."

Academic research has had mixed results regarding user preference with regards to male or female voiced computer systems. In a study comparing multiple text-to-speech (TTS) systems on measures including intelligibility and preference, Stevens et al. (2005) found that in two of the three TTS systems that were tested, female voices were rated as being more intelligible. Voice gender did not however have an impact on user preference. Nass et al. (2003) found that subjects were more likely to disclose personal

information to a female-voiced system. Dykema et al. (2012), with a sample comprised of young adults/adolescents, interviewed subjects utilizing a pre-recorded questionnaire. These researchers found that male subjects in the female voice condition disclosed sensitive information to a greater degree than male subjects in the male voice condition. This effect was not observed among female subjects. Evans and Kortum (2010) found no effect of voice gender on disclosure among users of a medical interactive voice response system.

The above sample of results seems to indicate a lack of consensus as it regards voice gender of computer systems and suggests that any effects if present might depend on a system's implementation and indeed the circumstances of its intended use. The current study sought to add to the literature regarding the potential impact of voice gender of a computer system by examining it in yet another context and by including performance in addition to preference as an outcome measure.

1.6. The Experiment

The current research sought to build on my previous work (Thomas & Lane, 2010) that examined the effect of the politeness of computer feedback through the application of social and cognitive theories. In this study, subjects completed a multiple cue probability learning task (MCPL) with either direct feedback or polite feedback. One interesting finding was that women improved their performance in both the direct and polite feedback conditions whereas men only improved in the polite condition. When asked which interface they liked the most, a majority of women selected one with polite feedback whereas men were more likely to select one with direct feedback. In that study,

61% (17/28) of women preferred polite feedback compared to only 26% (5/19) of the men. Figure 1 contains a graphical representation of these values. This difference was statistically significant χ^2 (1, N = 47) = 5.38, p = .020.

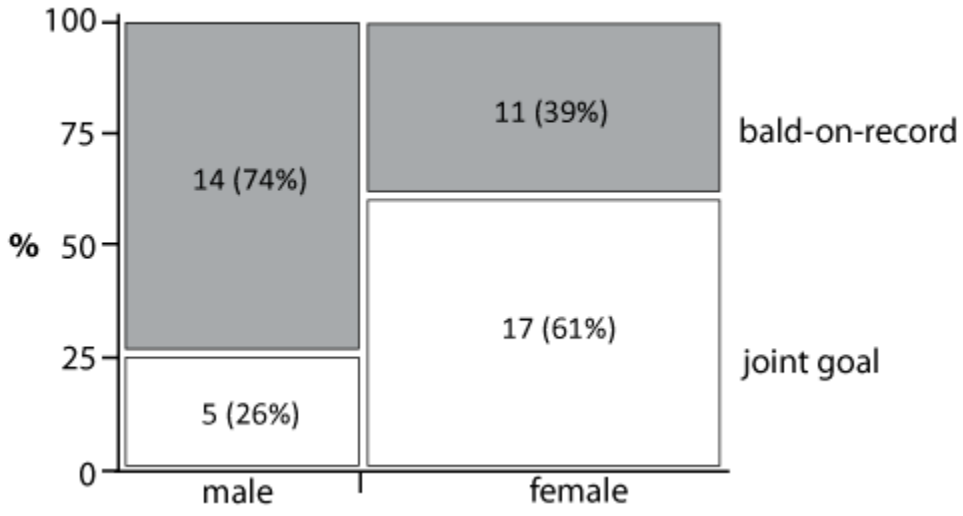

Figure 1 Gender by feedback tone (N = 47) ***This data is from Thomas and Lane (2010).***

These results suggest that men have a preference for the more direct type of feedback, whereas women tend to prefer less direct feedback, although not to the same degree. As a tentative explanation of this finding of a learning and preference effect between genders, Thomas and Lane (2010) make reference to research associated with social role theory. Specifically, Fagot (1985) in a study of very young children (21-25 months) found that boys did not respond to positive or negative feedback given by teachers or girls, whereas girls responded to both types of feedback from teachers. Roberts (1991) used social role theory to assert that men may be less responsive to evaluative feedback because they tend to be socialized in a more combative peer group

and so are more likely to ignore critical feedback. Women on the other hand tend to be socialized in a more collectively-oriented peer group and are thus more receptive to feedback in general.

In addition to the difference in politeness strategies used by the joint-goal prompts and the more direct bald-on-record prompts, there was also a difference in the grammatical person in which the prompts were phrased which might have impacted the preference effect. More specifically, joint-goal prompts were phrased in the first-person plural while direct feedback prompts were phrased in the second-person. It is possible that males were less receptive to the use of the pronoun “we” in the joint-goal prompts rather than the use of the auxiliary verb “should” or the modifier “it looks like”. Several participants reported that they thought that the polite prompts, by using the first-person plural, seemed somewhat “strange" and perhaps this decreased the politeness effect. The inability to separate grammatical person from politeness was a limitation of this previous study. The idea for a further examination of grammatical person was also encourage in a personal communication with Ben Shneiderman who in a discussion regarding the politeness/directness of computer feedback prompts, pointed out that the personal pronouns of these prompts were different and that in his experience subjects picked up on these differences, specifically in unpublished research conducted by his lab.

The current study expanded on the previous one by examining the politeness effect of feedback on learning/performance and user preference when learning with a computer, as well as (1) how this effect is impacted by the grammatical-person of the feedback given by the computer, (2) the mode in which it was delivered (i.e. whether text feedback differs from narrated feedback, and (3) whether gendered synthesized speech

has a different effect than gendered natural speech). Possible interactions of politeness with gender and with cognitive ability were also of interest.

1.6.1. The Learning Task

A key consideration for this research was the choice of a learning task. Two criteria were essential in selecting a task: first, the task had to provide multiple opportunities for feedback, secondly, there had to be an objective way of measuring user performance. Given these criteria the MCPL task was chosen.

The MCPL paradigm is best known for its relationship to the Brunswik lens model which holds that individuals adapt to their environment by learning how to predict future events from proximal cues (Brunswick, 1955). Brunswick developed this model around the philosophy of probabilistic functionalism which, in addition to recognizing this ability to predict future outcomes based on current cues, also held that these cues typically predict the outcome with less than 100% accuracy. The model is essentially a multiple regression where the values of a series of input variables can be used to predict the value of an outcome variable.

The lens model has been studied in a wide range of circumstances and in a variety of configurations. The model has been tested with children and young adults (Deffenbacher & Hamm, 1972; Lafon, Chasseigne & Mullet, 2004), and older adults (Chasseigne, Mullet & Stewart, 1997; Chasseigne *et al.*, 1999). It has been studied with a varying number of cues and different types of relationships between the criterion and the cues. Thus it is fair to say that the lens model is reasonably well understood with regards to MCPL which makes the difficulty level of the task easier to manipulate thus allowing

for a greater degree of experimental control when assessing the impact of various types of feedback.

There are several advantages to using multiple cue probability learning for assessing the value of feedback. Experiments involving this model consist of a series of trials divided into multiple blocks. The achievement of judges can be measured within blocks or across them. It is also possible to monitor the judge's performance by comparing the weights of the cues based on the judge's responses to the weights with the actual criterion values. This allows feedback to be tailored to the cues that the judge is having the most difficulty weighting correctly. Studies that utilize MCPL tasks have used a variety of approaches to providing feedback. This feedback can be categorized into three groups: outcome feedback, task information feedback, and cognitive/process feedback (Karelaia & Hogarth, 2008). Outcome feedback was characterized by Todd and Hammond (1965) as "knowledge of results" and refers to a condition where a judge is shown the correct criterion value after each trial. To link this with the types of feedback discussed previously, outcome feedback would be most closely related to the corrective feedback. The distinction between task information feedback and cognitive feedback is somewhat blurred in the literature. Some researchers define task information feedback as information that is provided regarding the relationship between individual cues and the environmental values of the criterion while cognitive feedback is seen as information regarding the relationship between the individual cues and the criterion values provide by the judge (Karelaia & Hogarth, 2008). Others do not make this distinction and instead combine the two by viewing cognitive feedback as information about the relationships both environmental and for the judge (see Balzer, Doherty, & O'Connor, 1989; Todd &

Hammond, 1965). Cognitive feedback is most closely related to process level feedback in the Hattie and Timperley (2007) model.

As will become clear, the only distinction that was important for this research is the one between outcome feedback and the other two, and as such task and cognitive feedback were both be referred to as cognitive feedback. Cognitive feedback is generally viewed as better for learning than is outcome feedback. Karelaia and Hogarth (2008) in a meta-analysis of lens model studies found that judges benefit more from information about the task rather than information about each trial. Todd and Hammond (1965) point out that while outcome feedback may be appropriate for simple learning tasks it becomes less helpful when it requires subjects to have to associate individual responses with individual cue configurations over a large number of trials.

1.7. Research questions

The literature raises several questions that this research attempted to address and these are as follows.

1. *How does the tone of formative feedback affect learning and preference of subjects?*
2. *Does feedback tone interact with the grammatical person that it is phrased in?*
3. *Are there differences in the politeness effect when feedback is narrated by a human voice compared to a computer- generated voice?*
4. *Are there differences in the politeness effect when feedback is narrated by a female voice versus a male voice?*

5. *Are their gender interactions with feedback tone, grammatical person and narrated feedback?*
6. *Is there an interaction between cognitive ability and the effect of politeness?*

How does the tone of formative feedback affect learning and satisfaction of subjects?

Affective computing research has demonstrated that users in their interactions with computers often treat these computers as social actors. Teachers can use varying degrees of politeness when giving feedback to students to influence how these students experience and accept this feedback. This research sought to further explore whether these politeness strategies can be applied when the teacher is replaced with a computer.

Does feedback tone interact with the grammatical person that it is phrased in?

There is evidence that people are not necessarily as comfortable with computer feedback that is phrased in the first person (Shneiderman et al., 2009). If subjects do not consciously believe that the computer is human then computer feedback phrased in the first person may be less effective than feedback phrased in the second person. There is somewhat of a tradeoff that has to be made between politeness and grammatical person as statements in the first person are generally thought to be more polite than those in the second (Mayer et al., 2006). However Mayer et al. (2006) also found that individuals are able to detect different levels of politeness between prompts phrased in the second person, which this study utilized.

Are there differences in the politeness effect when feedback is narrated by a human voice compared to a computer-generated voice?

Politeness is employed in interactions that social actors have with each other and while these exchanges can certainly occur with written text, they are often conducted verbally. It is of some interest therefore whether the politeness effect is affected by whether feedback is narrated and, furthermore, by the voice of the narrator. Of particular interest here is the difference between computer generated and human voices. On the one hand, it might feel more natural for a human voice to be polite than for a computer-generated voice. On the other hand, if the clearly inanimate computer appears to be "trying" to be too human, subjects may react negatively. A corollary of this type of reaction is the “uncanny valley” effect that can occur when viewing badly animated avatars.

The research utilized two types of male and female voice prompts each of which were patterned on a Caucasian, American accent: a human voice and a computer-generated voice. It is clearly beyond the scope of this research to test the whole space of voice types and I decided to increase the chance of finding an effect (if, indeed there is one) by choosing voices that differed greatly on how robotic they were. Thus one set of voices were relatively robotic and computer-generated and the other set relatively expressive and human.

Are there differences in the politeness effect when feedback is narrated by a female voice versus a male voice?

An issue that arises when using narrated text is whether the voice should be male or female. There is some evidence that a voice's gender does not matter. Evans and Kortum (2010) found no difference in how participants rated or disclosed information to male and female voiced medical interactive voice response (IVR) system. In general voices for computer systems tend to be female, with the most recent example being Siri, Apple's voice for the recently released iPhone 4s. A CNN article quotes Clifford Nass (a noted researcher in affective computing) "Its much easier to find a female voice that everyone likes than a male voice…" (cited by Griggs, 2011). This research further examined impacted of the gender of voice prompts.

Are their gender differences in the relationships between feedback tone, grammatical person and narrated feedback?

Given the limitations of the previously discussed research, this study attempted to further explore the gender by politeness interaction and unconfound politeness and gender. The research employed a more nuanced approach to measuring dominance and submissiveness by administering the Bem sex role inventory.

Is there an interaction between cognitive ability and the effect of politeness?

Mclaren, Deleeuw and Mayer (2011) found that the politeness effect was stronger for subjects with lower chemistry experience when learning with a computer application. Wang et al. (2008) found that subjects with less computer experience were more satisfied when learning via a polite computer application. Given the difficulty of the MCPL task it is plausible that cognitive ability may interact with the politeness of feedback.

Chapter 2

Method

2.1. Design

This study employed a mixed-factor design: Politeness/Person (polite first-person plural, polite second-person, and direct) x Mode (visual text, synthesized male voice, synthesized female voice, human male voice, human female voice) x Gender. Politeness/Person was a within-subject factor whereas mode was a between-subjects factor. For mode, subjects were assigned to five groups as shown in Table 2.

Mode of feedback	
visual text	56
synthesized male voice	23
synthesized female voice	23
human male voice	24
human female voice	24

Table 2 Number of subjects assigned to each mode

The questions posed by this research were characterized by the two primary outcome measures: preference and learning. Preference was measured by subjects' selection of the interface they felt had the best feedback at the end of the experiment as well as their rating of the feedback at the end of each condition. Learning was measured both in terms of the extent to which subjects were able to reduce the difference between their predictions and those of the model over blocks of trials during the MCPL task and also the ability to identify the most and least heavily weighted test in each of the three feedback conditions.

A concern that is often associated with within-subjects designs is whether subjects notice the manipulation as they progress through the conditions of the study. Placed in the context of this study, the concern is whether subjects overtly noticed changes in tone and grammatical person of the feedback they receive in a way that affects their responses. To account for this possibility, subjects were asked if they noticed any differences in the tone of feedback at the end of the experiment. This variable is included in the analyses.

2.2. Subjects

A total of 150 subjects recruited from the Rice undergraduate student population participated in this study. All subjects received course credit for their participation. Two subjects were excluded from the final analysis. The first of these subjects recorded a very high median score of 203 for their final block of learning trials in the joint-goal condition This score was 11.67 standard deviations away from the mean judgment error in the joint-goal condition. Higher values in judgment error represented a greater disparity in

the subject's predictions compared to the model's prediction in the MCPL task. The second subject to be excluded did not participate in one of the feedback tone conditions.

The final sample of 148 subjects consisted of 69 males and 79 females with a mean age of 19.5 years and an age range of 18 years to 37 years).

2.3. Materials

The MCPL task was programmed using HTML, JavaScript and PHP. Figure 2 through Figure 4 are screenshots of the interfaces used in the experiment. The application was run in Chrome on a Macintosh computer. The Interfaces differed visually, but provided identical interaction schemes with the MCPL task, in that subjects could advance a trial by typing their prediction and then pressing the "Enter" key.

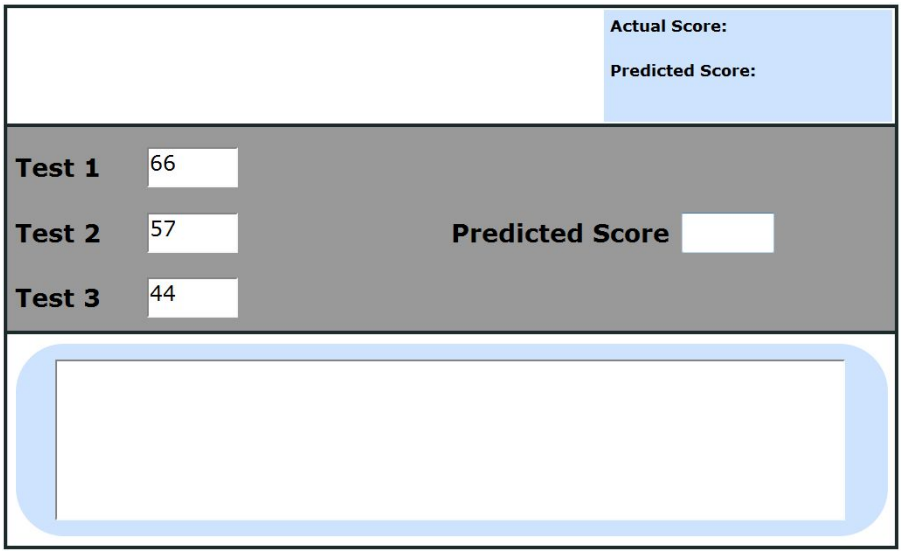

Figure 2 Interface 1

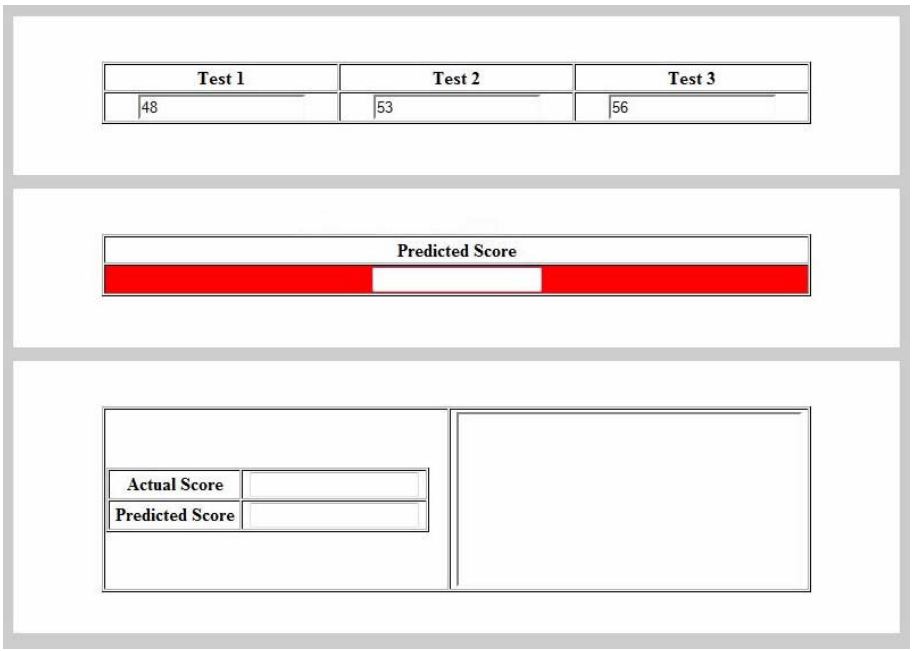

Figure 3 Interface 2

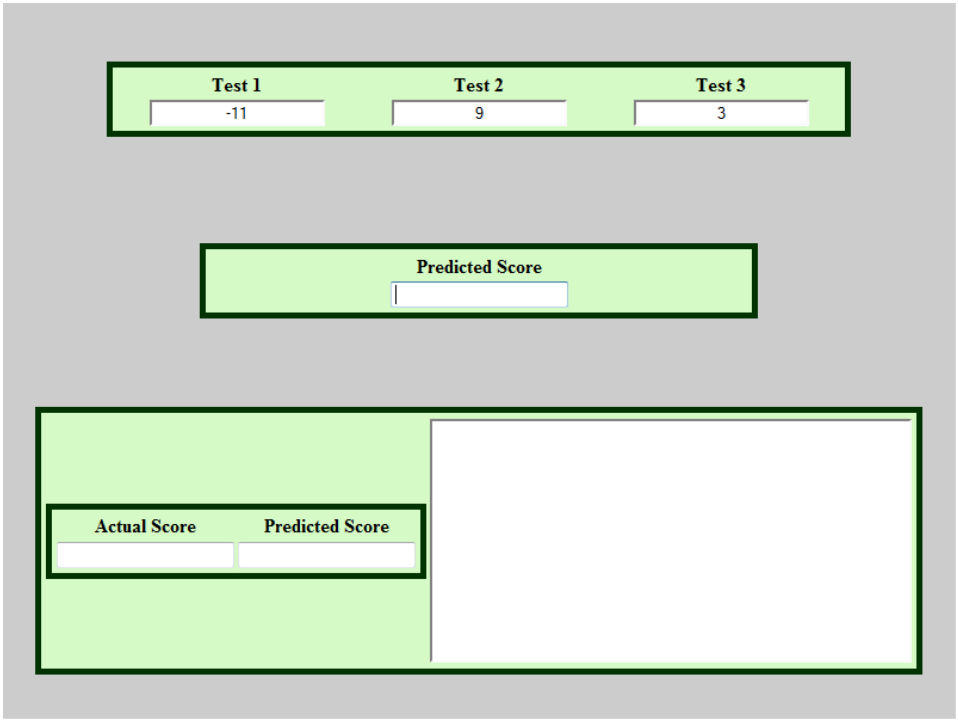

Figure 4 Interface 3

MCPL stimuli were generated using the R statistical package and were structured such that there were three orthogonal cues with cue validities of 2, 1 and 0 (raw weights). These validities varied slightly for the actual experiment as all values were rounded to integers to reduce the task load on subjects by not requiring them to have to estimate fractional values. In an additional effort to simplify the task, environmental predictability was high and ranged from .97 - .98. The mean for each cue was set to zero. Figure 5 is a representation of the MCPL learning task where Y_e is the criterion variable in the environment and Y_j represents the judgments of the criterion variable. X_1 to X_3 represent the predictors the weights of which are displayed on the connections with Y_e.

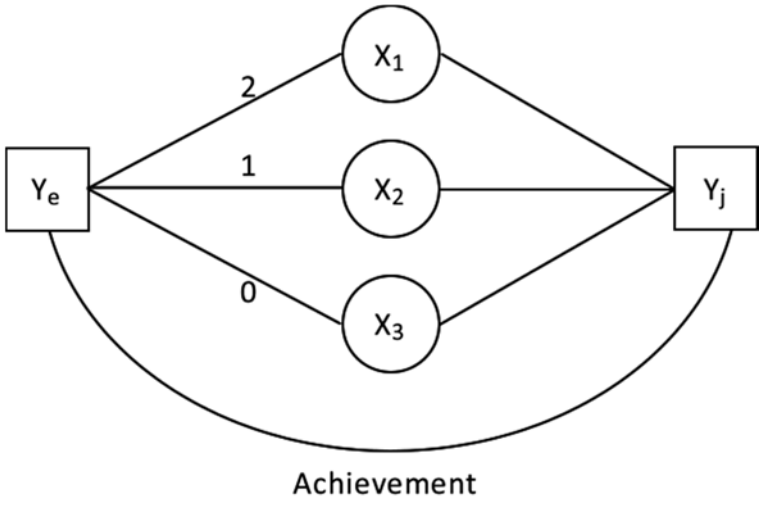

Figure 5 Model of the MCPL learning task used in experiment.

Table 3 shows examples of the feedback prompts that were used in the MCPL task.

Prompt Type		Sample Statement
bald-on-record	second person	The scores you entered highly over weigh the impact of test 2, you must focus more on the other tests.
student-goal	second person	It looks like the scores you entered still highly over weigh the impact of test 2, you should pay more attention to the other tests.
joint-goal	first person plural	It looks like the scores we entered highly under weigh the impact of test 1, we should pay more attention to this test.

Table 3 Feedback prompts in experiment

This study used several personality and cognitive ability measures to differentiate between subjects.

2.3.1. Dominant versus Submissive personalities

The Bem Sex role inventory (BSRI) is a sixty item scale that was originally proposed by Bem (1974) as a measure of masculinity, femininity and androgyny. Subjects rate themselves on a seven point Likert scale on each item, their responses are aggregated and both a masculinity score and femininity score are calculated. While this measure has been fairly popular it has been criticized with regards to its construct validity (Ballard-Reisch & Elton, 1992; Choi & Fuqua, 2003). Its detractors point out that gender roles have changed since the mid-70s and that therefore the items used in the BSRI to make the masculine versus feminine distinction no longer applicable.

Instead of dismissing the measure however, some research has shown that subscales of the BSRI correlate very highly ($.8 \leq r \leq .87$) with other measures of dominance and agreeableness (Wiggins & Holzmuller, 1981; Wiggins & Pincus, 1989). These correlations are achieved with the removal of a few items that were either not relevant conceptually, or were found to be unreliable. Wiggins & Pincus (1989) regard the use of the labels masculine and feminine as "unfortunate" and suggest that the labels of ambitious-dominant and warm-agreeable be used instead. Table 4 displays the items of the BSRI divided into their subscales and indicates which items were removed for conceptual or reliability reasons to achieve the above correlations.

Feminine	Masculine	Neutral
2. yielding	1. self-reliant	3. helpful
5. cheerful	4. defends own beliefs	6. moody
8. shy	7. independent	9. conscientious
11. affectionate	**10. athletic**	12. theatrical
14. flatterable	13. assertive	15. happy
17. loyal	16. strong personality	18. unpredictable
20. feminine	19. forceful	21. reliable
23. sympathetic	22. analytical	24. jealous
26. sensitive to the needs of others	25. has leadership abilities	27. truthful
29. understanding	28. willing to take risks	30. secretive
32. compassionate	31. makes decisions easily	33. sincere
35. eager to soothe hurt feelings	34. self-sufficient	36. conceited
38. soft-spoken	37. dominant	39. likable
41. warm	**40. masculine**	42. solemn
44. tender	43. willing to take a stand	45. friendly
47. gullible	46. aggressive	48. inefficient
50. childlike	49. acts as a leader	51. adaptable
53. does not use harsh language	**52. individualistic**	54. unsystematic
56. loves children	55. competitive	57. tactful
59. gentle	58. ambitious	60. conventional

Table 4 Items in the BSRI. Non-bolded items lead to high correlations with warm-agreeableness factors for the feminine subscale and ambitious-dominant factors for the masculine subscale.

Nass et al. (1995) used the adjusted masculinity scale to distinguish between dominant and submissive personalities in their subject pool. On the basis of this theoretical and practical evidence, the BSRI was used for this study as well.

2.3.2. Cognitive ability and math anxiety

This research used self-reported SAT scores as a measure of cognitive ability. In an oft cited article, Frey and Determan (2004) found a correlation of .82 between SAT scores and measures of *g*. There has been general concern about the accuracy of self-reported SAT scores in the literature, in a meta-analysis Kuncel, Credé and Thomas (2005) reported a correlation of .82 between overall self-reported and actual SAT scores. The paper also reported that the correlation between the self-reported SAT verbal (r = .74) and mathematical (r = .82) and their respective actual SAT sub-test scores.

Of the 150 subjects, 130 reported their overall SAT scores and 14 reported ACT scores. ACT scores were converted to their SAT equivalents. Four subjects reported scores from the 1600 SAT, these scores were converted to the 2400 scale. 138 subjects also reported their SAT/ACT math scores. Table 5 displays descriptive statistics of both SAT scores. The mean overall and math SAT scores were both fairly high for the sample at 2,140.69 and 735.18 respectively.

	Mean	SD
Overall SAT	2,140.69	142.33
Math SAT	735.18	60.74

Table 5 Mean and standard deviation of SAT scores

Subjects were asked to complete the Math Anxiety Scale-Revised (MAS-R). The 14 item MAS-R, developed by Bai et al. (2009) was adapted from an instrument that was designed to assess math anxiety among college students (Betz, 1978). Each item is rated

on a 5-point scale with 1 labeled as “strongly disagree” and 5 “strongly agree.” See Table 6 for a list of items in the scale.

Items 2,4,6,7,8,9,11,14 are negatively phrased and were reverse coded. A final anxiety score was calculated for each subject by adding their responses to the fourteen items together. Scores can range from 14 to 70 with higher scores representing lower math anxiety. The mean MAS-R score for the sample was 48.41 with a standard deviation of 10.39 with scores ranging from 17 to 69.

1	I find math interesting.
2	I get uptight during math tests.
3	I think that I will use math in the future.
4	Mind goes blank and I am unable to think clearly when doing my math test.
5	Math relates to my life.
6	I worry about my ability to solve math problems.
7	I get a sinking feeling when I try to do math problems.
8	I find math challenging.
9	Mathematics makes me feel nervous.
10	I would like to take more math classes.
11	Mathematics makes me feel uneasy.
12	Math is one of my favorite subjects.
13	I enjoy learning with mathematics.
14	Mathematics makes me feel confused.

Table 6 Items in the MAS-R

2.3.3. Voice Prompts

Four voices were used for the audio feedback prompts in this study. The human voices, one male and the other female, were sampled from two Rice students who both spoke with an American, Caucasian accent. Both voice actors were given fairly simple instructions in their reading of the prompts and asked to speak in a clear and expressive manner. The International Dialects of English Archive (IDEA) is a publicly accessible online database of English dialect recordings and the IOWA1[1] sample in this collection is a reasonable approximation of the human female voice used in this study. The Connecticut 1[2] sample is a good approximation of the male voice used. Synthesized voice prompts were captured with the Macintosh text-to-speech (TTS) engine. The criteria for the synthesized voices were that they spoke in a stilted, robotic style that would allow them to be easily distinguishable from human voices while still maintaining a level of clarity that was comparable to human voices.

Four voices (two male and two female) that mimicked American accents were selected and these, along with the human voices were evaluated by three raters on three dimensions: roboticness, clarity and listening effort. Each of these dimensions was measured on a 7 point Likert scale that raters completed for each voice after listening to an audio sample. Figure 6 summarizes the results of the ratings for the three raters.

---

[1] http://www.dialectsarchive.com/iowa-1 (IDEA)
[2] http://www.dialectsarchive.com/connecticut-1 (IDEA)

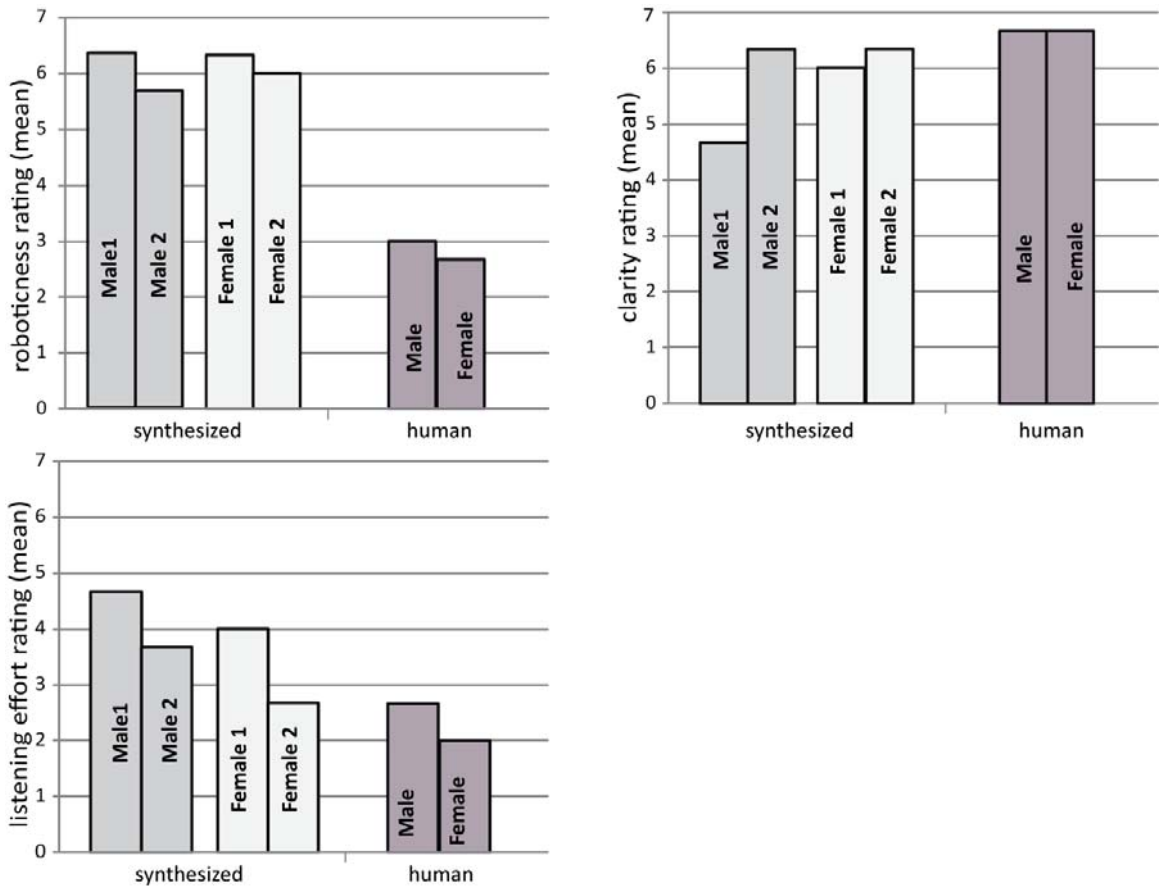

Figure 6 Mean ratings of audio prompts on the three dimensions: Roboticness, Clarity and Listening effort (N = 3)

The results show a clear distinction between human and synthesized voices on the roboticness dimension, while the male 2 and female 2 synthesized voices compared most favorably to the human voices on the listening effort and clarity dimensions. These two synthesized voices along with the human voices were the ones used for the voice prompts in this study.

2.4. Procedure

Subjects were asked to sign a consent form at the start of the experiment. Having agreed to participate, subjects completed the Bem Sex role inventory and a short demographic questionnaire that included a question about their SAT scores. Upon completion, subjects were read a scripted statement providing an overview of what the experiment would require them to do. This statement varied depending on whether the subject had been assigned the narrated text or visual text condition. Before beginning the MCPL task, the experimenter made sure subjects had a reasonable understanding of how to approach the task by asking them to predict the value of a criterion variable given the weight and value of a single cue. Subjects did not proceed to the task until they could successfully make this prediction.

The MCPL task was framed within the context of predicting high school students' second year performance based on their score on three first year tests. Subjects were given instructions that their task was to discover how important each of the first year tests were in making a prediction. It was stressed that this was a training application and that the purpose of the exercise was to evaluate three differing interfaces. It was important to not have subjects discover that the focus of the experiment was on the politeness/directness/grammatical person of the feedback prompts as this could have biased their learning outcomes and subjective ratings.

Subjects completed 80 trials in each of the three politeness/grammatical person conditions with feedback either as visual text, a synthesized voice or a human voice with a different interface used in each. The order of the interfaces and the conditions were

counterbalanced across subjects. At the beginning of every trial, a subject was shown three first year test scores and asked to predict a second year score. After submitting a response, subject were shown what the actual score was (outcome feedback). The raw *b* weights of the cues remained constant for all subjects in all three feedback tone conditions so the three cue weights were always 2, 1 and 0. However, the order of the cues was randomized across subjects and conditions, in other words for any given set of 80 trials Test 1 may have had any of the three cue validities.

After every ten trials, the application presented a summative feedback prompt regarding the subject's performance over the preceding block of trials (cognitive feedback). These prompts varied in grammatical person, politeness and mode depending on the condition that the subject was in. Feedback statements were based on a comparison of the real weights compared to those calculated from the subject's predictions. The displayed/narrated statements provide cognitive feedback about which cue weight (calculated based on the subject's scores) was the most different from its real cue weight. In addition to stating which cue estimate was the worst, the application also displayed the real cue validity as well as the cue validity generated from the criterion values that were entered.

Trials were self-paced, but subjects were advised to not spend an excessive amount of time (more than four or five seconds) thinking about each of their response and to rely on their first impression. Each interface had a visual cue that indicated when five seconds had elapsed. Subjects were instructed to use this indicator as a guide and not as a time limit. Following each session of 80 trials, subjects rated the interface they used and indicated which tests they believed to be most and least heavily weighted. After

completing all conditions of the task, subjects were asked to indicate which interface they found to be more helpful and which they found to be more aesthetically pleasing.

Chapter 3

Results and Discussion

As discussed in the Design section of the Methods chapter, subjects were asked whether or not they noticed differences in the phrasing of feedback across the different interfaces after completing the three feedback conditions. This variable which will from here on be referred to as "noticed-phrasing" will be included in these analyses. The inclusion of noticed-phrasing in the analysis of user preference is not difficult to justify. When responding to the question of which interface had the best feedback, it was desirable to have subjects respond based on some subconscious impression rather than an overt recognition of how feedback varied across the three conditions, because this could potentially bias their response. It is somewhat more difficult to see how this recognition could impact learning however, but one could speculate that being aware of the differences in phrasing could cause a subject to attend to feedback in a way that makes them less susceptible to its tone which could impact learning.

Of the 131 subjects that responded to this question 78 (59.5%) reported that they did not notice the differences in the phrasing of feedback.

3.1. Feedback Preference

Subjects were asked to indicate which interface they felt provided the best feedback. Figure 7 displays preferences as a function of tone, gender, and noticed-phrasing. Among subjects that noticed differences in phrasing, the distribution of preferences was more or less equal across the three different types of feedback with the exception to this being men, who were somewhat more likely to select the more polite (joint-goal) feedback. For subjects that reported not noticing differences in phrasing, the proportions in the three feedback groups were not as uniformly distributed. Men were more likely to prefer an interface with feedback that was phrased as student-goal (46.7%) compared to only 12% of women.

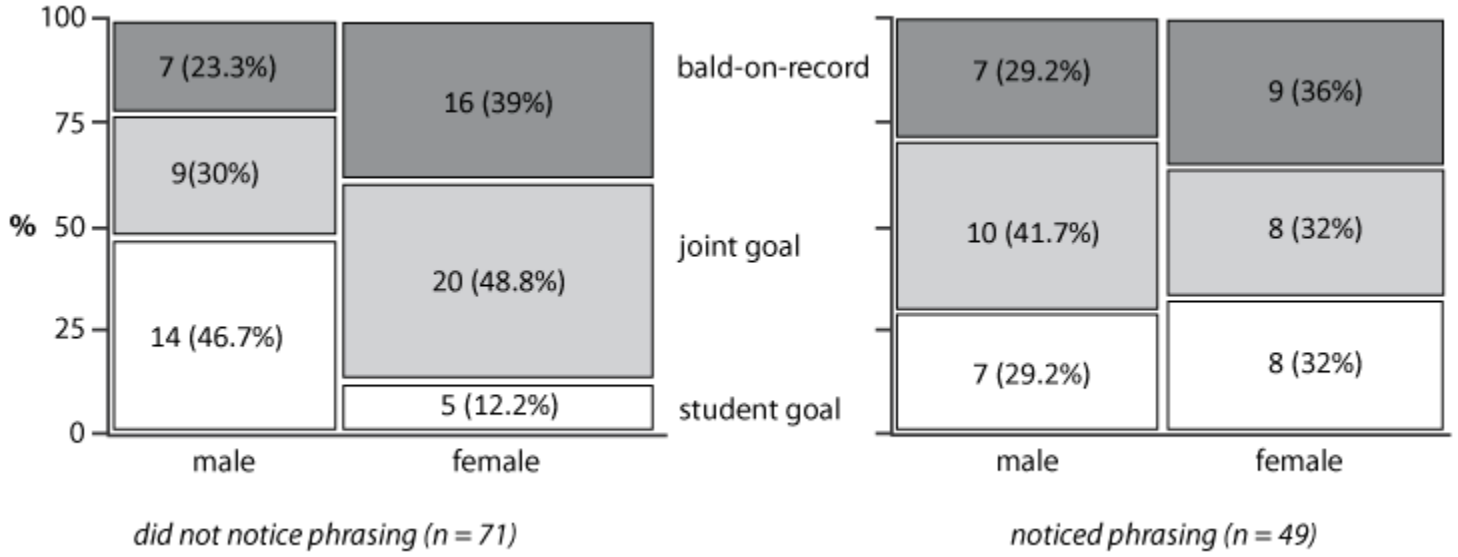

Figure 7 Proportion of subjects that preferred specific feedback by tone, gender, and notice

A multinomial logistic regression model including preference as the dependent variable and gender and noticed-phrasing as factors revealed suggestive but not conclusive evidence of a Tone x Gender x Noticed-Phrasing interaction , $\chi^2(2, N = 120) = 5.67$, $p = .059$, $C = 0.22$. Although this model did not achieve a conventional level of statistically significance, I felt that, given the effect size and the variability in proportions particularly among subjects that reported not noticing the differences in phrasing that it was worth exploring the simple effects in this model using conservative statistical methods.

The Gender x Tone simple interaction among those that did not notice phrasing, was significant $\chi^2(2, N = 71) = 10.624$, $p = .005$, $C = 0.359$. Follow-up simple effect comparisons of males to females for each feedback condition among those that did not notice differences in phrasing are shown in Table 7.

joint-goal	$\chi^2(2, N = 71) = 2.53$, $p = .112$, $C = 0.185$
student-goal	$\chi^2(2, N = 71) = 10.50$, $p = .001$, $C = 0.359$
bald-on-record	$\chi^2(2, N = 71) = 1.95$, $p = .163$, $C = 0.163$

Table 7 Simple effect comparisons of males to females tone preference among subjects that did not notice differences in phrasing.

Applying the Bonferonni correction because of the number of effects analyzed, (adjusted $\alpha = .017$) the gender effect was significant for student-goal feedback but not for the other types of feedback. A similar analysis among subjects that noticed differences in phrasing was not significant $\chi^2(2, N = 49) = 14.136$, $p = .771$, $C = 0.288$.

This analysis was repeated with the preference variable regrouped based on grammatical person of the feedback i.e. student-goal and bald-on-record grouped as second-person singular (you) compared to joint-goal (we). No statistical significant results were obtained with the Gender x Notice-Phrasing interaction yielding the following $\chi^2(1, N = 120) = 2.45$, $p = .118$, $C = 0.14$. Models including mode of feedback, math anxiety, SAT math score and the dominance/submissive personality scores were evaluated for both categorizations of the preference variable, however none of these variables resulted in consistent or significant patterns.

So what do these results indicate? Focusing primarily on subjects who did not discern the experimental manipulation of feedback during the learning task, men appeared to have a more positive view of student-goal style prompts than women who in turn were more likely to prefer joint-goal prompts as well as bald-on-record prompts. The student-goal politeness strategy, by addressing the receiver directly in the second-person as is the case with the bald-on-record strategy, while employing modifiers such as "should" as in the joint-goal strategy, can be seen as a middle ground between these two strategies.

Although the pattern of preferences of the two genders does not entirely replicate that of the previous study, there a few consistencies worth noting. Although not statistically significant, the proportion of men and women that preferred joint-goal style feedback was similar in both studies. This pattern of the two genders is even more clearly seen among subjects who reported not noticing a difference in phrasing. The same cannot be said for preferences of bald-on-record feedback. Whereas men were more likely to

prefer bald-on-record style feedback when compared to women in the previous study, this pattern was reversed in the current study, including for those who did not notice feedback phrasing.

There are a number of possible explanations for the discrepancies between this study and the results in Thomas and Lane (2010). One possibility is that the feedback category student-goal was not present in the first study. Instead of having to choose between two levels of feedback, subjects had a third option, which in the case of male subjects became the preferred choice. This suggests that men may not have a preference for direct feedback as a whole, but may indeed be less receptive to use of the first person plural in feedback prompts. Although there is no way of knowing what subjects that preferred the student-goal style feedback would have chosen were it not an option, I do not believe that it is unreasonable to speculate that the male and female subjects might select one of the other two types of feedback in line with male and female subjects in Thomas and Lane (2010).

3.2. Learning/Performance

Learning was approximately equal and not significantly different across all three types of feedback for subjects as a whole. Table 8 displays the means and standard error for both the absolute difference between the MCPL model and subjects predictions (which will from here on be referred to as judgment error) as well as the accuracy with which subjects were able to select the least and most important tests for each type of feedback. It should be noted that achievement, measured as the correlation between the

judge's and the model's scores, is a more traditional measure of performance in studies that use the lens model. Thomas and Lane (2010) found the correlation between judgment error and achievement to be very high yielding coefficients in excess of .95. Given this it was felt that it would be appropriate to use judgment error as a measure of performance for this study as differences scores in errors tend to be a more typical measure of performance in studies of this nature.

	judgment error		least important test		most important test	
	mean	std. error	mean	std. error	mean	std. error
joint goal	5.05	0.46	0.86	0.04	0.95	0.02
student goal	4.89	0.41	0.96	0.02	0.98	0.02
bald-on-record	4.70	0.46	0.89	0.03	0.92	0.03

Table 8 Means and standard errors of learning measures by tone

Subjects were good at selecting both the least and most important tests at the end of the trials in each condition. They were however, better at selecting the most important test, $F(2, 110) = 8.41$, $p = .005$. Judgment error was practically the same across the three tones. Subjects were slightly less accurate in the joint goal condition, but this difference was not statistically significant $F(2, 220) = 0.59$, $p = .560$.

Subjects were able to reduce judgment error across blocks of trials in all feedback tone conditions and at a similar rate as illustrated in Figure 8. Also clear from Figure 8 is that the tone by block interaction is essentially nonexistent and clearly not statistically significant $F(14, 1554) = 0.72$, $p = .753$.

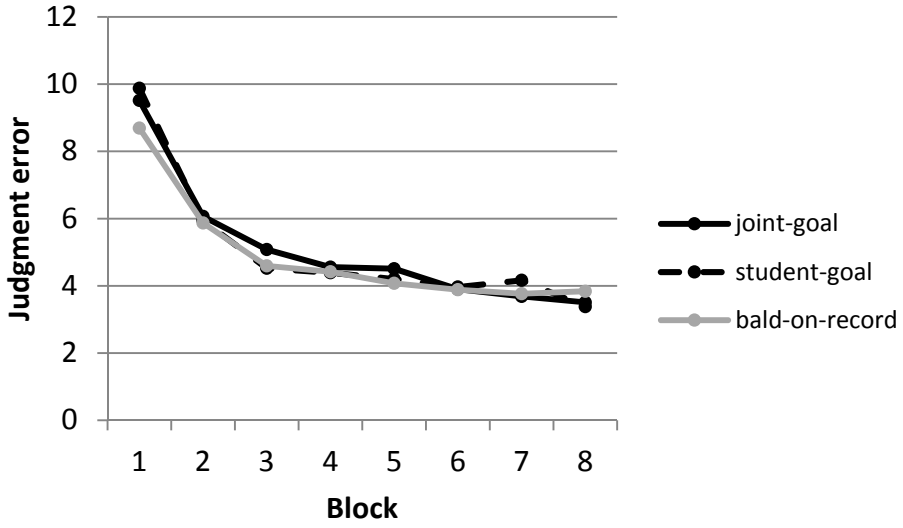

Figure 8 Judgment error by tone and block

This rate of learning over blocks of trials was similar across all measures in this study as the various Block x Tone interactions were found not to be statistically significant. Table 9 summarizes these results. A statistically significant result was found for the Tone x Block x Gender x Notice interaction ($F(14, 1554) = 1.74$, $p = .042$). Further exploration found that the 7th order contrast of this interaction was significant; however a graphical examination of this contrast uncovered no meaningful pattern.

Tone x Block x Gender	$F(14, 1554) = 0.955$, $p = .498$
Tone x Block x Mode	$F(56, 1554) = 1.05$, $p = .343$
Tone x Block x Gender x Noticed	$F(14, 1554) = 1.74$, $p = .042$
Tone x Block x Mode x Noticed	$F(56, 1554) = 1.31$, $p = .062$

Table 9 Statistics of learning (Tone x Block) with assorted factors

As can be seen in Table 10, judgment error was correlated with cognitive ability and math anxiety.

	sat	math sat	masr
judgment error	-.278, $p = .001$	-.232, $p = .006$	.35, $p < .001$

Table 10 Correlations of judgment error with cognitive ability and math anxiety

The learning measure correlated with each cognitive ability measure in the expected way, negatively with the SAT scores and positively with the math anxiety scores. These results provide a modicum of criterion validity as those subjects that reported higher quantitative and general SAT scores as well as those that reported lower math anxiety also tended to preformed better at this highly quantitative task.

Although there was no main effect of tone, the Tone x Noticed-Phrasing interaction yielded an interesting result. Figure 9 displays the Tone x Noticed-Phrasing interaction for judgment error. There were greater differences among conditions for those who did not notice phrasing. Specifically, those with the joint-goal condition had the highest error rate and bald-on-record the lowest. The Tone x Noticed-Phrasing interaction was statistically significant $F(2, 222) = 3.35$, $p = .037$. For the group that did not notice, the difference between the joint-goal and bald-on record was significant $t(77) = 2.10$, $p = .039$. These results should be interpreted with some caution since this difference would not have been significant if the Bonferroni correction had been used. With this in mind, this result suggests, but by no means demonstrates an impact of tone on learning among subjects who were not overtly aware of differences in phrasing in the three tone

conditions. For these subjects, judgment error was highest in the polite condition and least in the most direct condition. An explanation for this can be made by utilizing politeness theory which points out that joint-goal prompts while being more polite also tend to be less efficient leading to longer prompts. It is conceivable that the shorter, more efficient bald-on-record prompts were easier for subjects to attend to and process leading to improved performance.

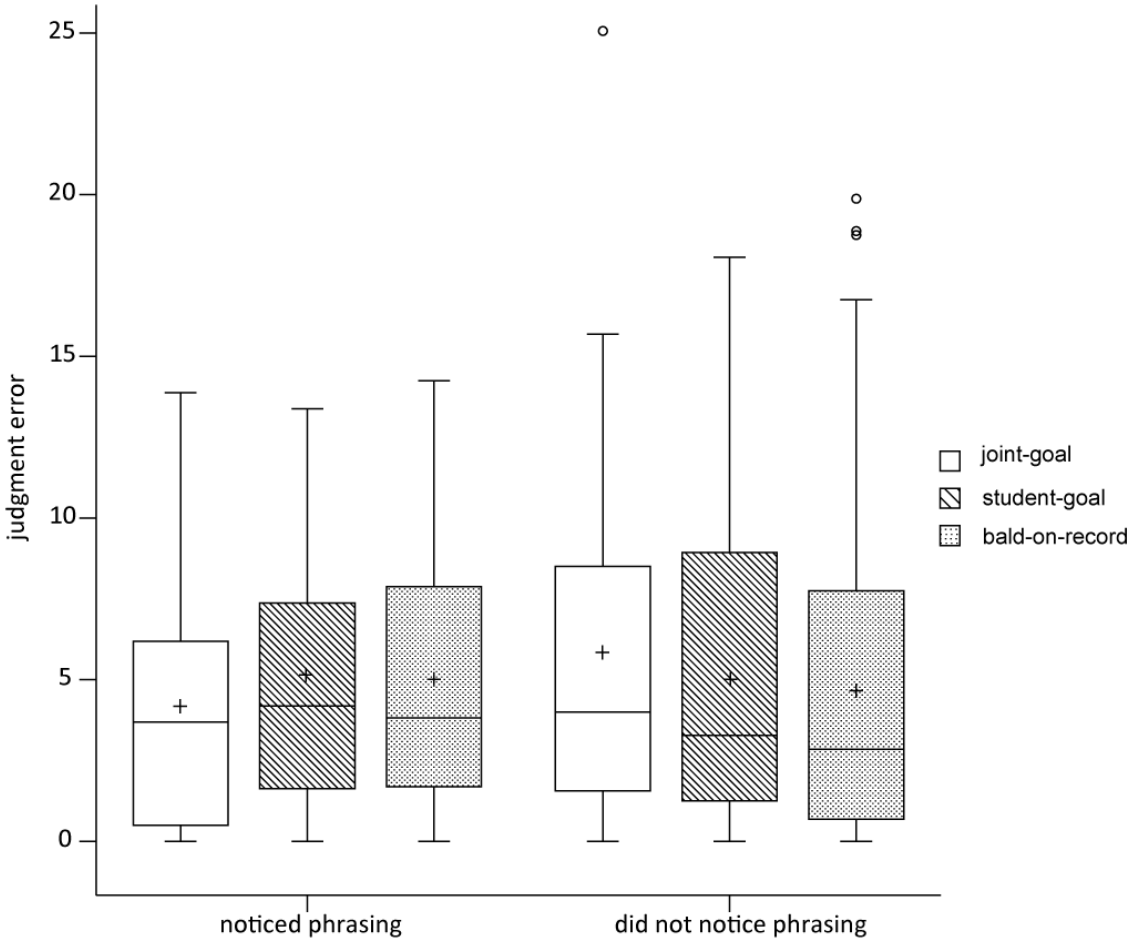

Figure 9 Judgment error by Tone and Notice-Phrasing

An important feature of this study was the inclusion of audio feedback, which allowed for a comparison of how subjects responded to text-based feedback and verbal feedback delivered as gendered synthesized or human speech. Figure 10 displays judgment error by tone, mode and noticed-phrasing.

There were little to no differences among feedback modes with the following two exceptions: (1) For those who noticed the phrasing and heard the synthesized male voice, subjects were much less accurate when receiving joint-goal feedback than when receiving either the student goal or bald on record feedback and (2) for those who did not notice the phrasing and heard the human female voice there were more errors made with joint-goal feedback than with student-goal feedback which in turn, had a higher error rate than those receiving bald on record feedback. There is a caveat in analyzing this result, since both mode and notice-phrasing were between-subjects variables it is important to keep in mind the number of subjects in each group and to analyze the Notice x Mode x Tone interaction with caution. This was particularly true for the synthetic male voice condition among those that notice phrasing which was comprised of only five subjects.

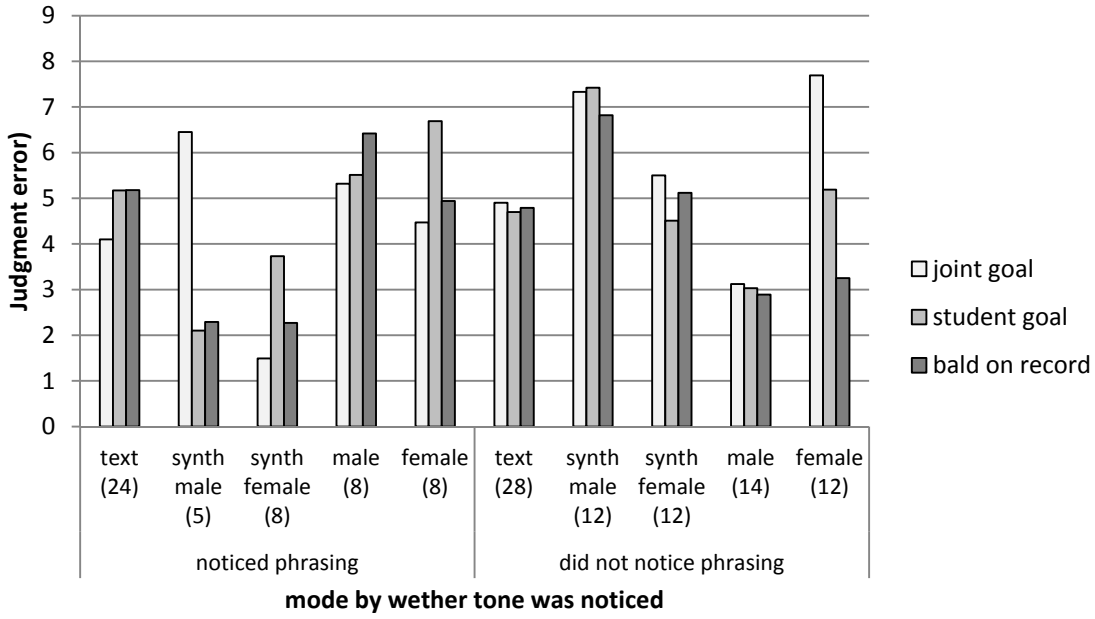

Figure 10 Judgment error by feedback mode by noticed. Parentheses on x-axis display n for each group.

The Notice x Mode x Tone interaction was statistically significant $F(8, 222) = 2.53$, $p = .019$. The Mode x Tone interaction was significant both for those who noticed the phrasing, $F(8, 86) = 2.13$, $p = .042$, and for those that did not notice the phrasing, $F(8, 136) = 2.08$, $p = .042$. The main effect of Tone in the human female voice condition among those that did not notice phrasing was also statistically significant $F(2, 22) = 4.76$, $p = .019$. Post hoc t-test comparisons were carried out for the feedback tone conditions within this group the results of which are displayed in Table 11.

Feedback tone pairs	
joint-goal — bald-on-record	t(11) = 2.71, p = .020
joint-goal — student-goal	t(11) = 1.40, p = .189
bald-on-record — student-goal	t(11) = 2.19, p = .051

Table 11 Post hoc comparisons feedback tone conditions in the female voice mode among for subjects that did not notice phrasing

Subjects receiving female voiced feedback achieved significantly lower judgment error in the bald-on-record condition than in the joint-goal condition. The difference between the other two comparisons did not attain conventional levels of statistical significance. It should be noted that the joint-goal to bald-on-record comparison would not be reliably significant when the Bonferonni correction is (adjusted $\alpha = .017$).

Additional analysis with judgment error and gender as well as measures cognitive ability and math anxiety were carried out but yielded no compelling interactions. Table 12 summarizes these results.

Tone x Math Anxiety	$F(2, 252) = 0.37$, $p = .690$
Tone x Gender x Noticed	$F(2, 252) = 1.05$, $p = .343$
Tone x SAT	$F(2, 248) = 0.93$, $p = .396$
Tone x Math SAT	$F(2, 240) = 0.08$, $p = .919$
Tone x Personality (dominance vs submissiveness)	$F(2, 256) = 1.86$, $p = .158$

Table 12 Statistics of judgment error with assorted factors

The quantitative measures math anxiety, SAT, Math SAT and personality were centered and used as covariates in the analysis presented in table 10.

3.3. Conclusion

The tone with which learning applications communicate with users appears to have an impact on user preference and possibly learning. Although some of the effects should be interpreted with some caution, they are consistent with previous findings which, gives credence to their reliability.

This study provides evidence that men are more likely to prefer student-goal style prompts whereas women are more likely to prefer one of the other two styles. This finding is particularly interesting when examined in conjunction with the preference finding from the previous study (Thomas & Lane, 2010) with the MCPL task. In this earlier work, men tended to prefer bald-on-record prompts over joint-goal prompts, with the reverse being true for women. This finding on its own suggested that men, when receiving feedback from a learning application preferred a more direct tone rather than one that was more polite. The finding in the current study allows for a more nuanced possibility in which men, rather than preferring a more direct tone, instead prefer a more polite tone that addresses them in the second-person.

The only significant learning effect associated with feedback tone found in this study was an interaction with feedback mode and whether or not subjects reported noticing differences in feedback tone between conditions. An examination of this somewhat complex three-way interaction found that among subjects who reported not noticing tonal differences and to whom feedback was delivered in a human female voice performance was better with bald-on-record style phrasing than with joint-goal phrasing. The mixed results in the literature regarding voice gender of computer systems (e.g. Nass

et al., 2003; Evans and Kortum, 2010) do not provide a particularly clear answer to the question of whether a computer's voice should be male or female. Although this result does not do much to clarify this question it does add an additional wrinkle, where female voices that are more polite/friendly could potentially have a negative impact user performance.

This research is certainly not without its limitations some of which I think it is important to point out. Perhaps the most apparent one of these is the homogeneity of the sample used. Although there was some variability in cognitive ability as evidenced by subjects' SAT scores and math anxiety, the fact that this was a college sample means that one should be cautious in generalizing these findings to the general population. In addition to this, it is probably safe to assume a reasonable degree of computer experience and aptitude in completing computer-based tasks across all subjects in the experiment. Individuals with less computer experience may well be more sensitive to the tone in which computer prompts are phrased, but this comparison was not possible with the sample used.

The structure of the learning task was one in which feedback was delivered at regular predefined intervals regardless of how a participant was proceeding from trial to trial. Although this type of structure is not unusual particular with regards to learning applications, it is also common for a teacher, whether human or virtual, to vary the frequency of feedback based on a learner's performance during the task. So instead of everyone receiving an equal amount of prompts, weaker learners would receive more prompts than stronger learners.

Several recommendations as well as directions for future research arise out of this work. Although this and previous studies have not found consistent effects of feedback tone on learning, there do seem to be reliable differences in user preference for specific tonal styles. This study found gender differences in preference for specific tones even for this fairly homogenous sample. On this basis one could conjecture that these differences and perhaps others might appear in more diverse populations. Repeating this study in the general population would of course require further adjustments to the MCPL learning task and perhaps require an entirely different task altogether. The MCPL task is ideal for manipulating the rate at which a subject receives feedback across trials, but given how difficult it was to balance its difficulty for this college sample, this will be even more difficult to do for the general population.

An interesting future direction would be to examine if users, in completing a learning task, are more likely to request feedback from a computer given a particular tonal style. In the current study, feedback was tied to the number of trials that a subject had completed however one could imagine a system where a subject could request feedback much in the same way that student could ask a teacher a question about his/her progress. If users are more likely to solicit aid from a computer that uses a particular style of feedback then this could help to inform the design and implementation of these systems.

The existence of social convention in human computers interactions has been demonstrated in a variety of domains. This study focused on the learning domain where the human-computer relationship could be characterized as a student-teacher relationship.

The politeness of the feedback prompts used by a learning application was found to have some impact on user preference, in particular when comparing male and female subjects. Feedback tone also had a minimal impact on learning/performance, with the only reliable effect occurring when prompts were delivered in a human female voice. Despite these results it seems worthwhile to explore whether these effects might exist in more diverse populations. Designers of learning applications should be aware that the genders may react differently to different types of feedback and successful designers will take this into account.

References

Baddeley, A. (1992). Working memory, *Science, 255,* 556 – 559.

Bai, H., Wang, L., Pan, W., & Frey, M. (2009). Measuring mathematics anxiety: Psychometric analysis of a bidimensional affective scale. *Journal of Instructional Psychology, 36,* 185-193.

Ballard-Reisch, D., & Elton, M. (1992). Gender orientation and the Bem sex role inventory: A psychological construct revisited. *Sex Roles, 27*, 291 – 306.

Balzer, W. K., Doherty, M. E., & O'Connor, R. Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*, 410 – 433.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*, 155 – 162.

Betz, N. E. (1978). Prevalence, distribution, and correlates of math anxiety in college students. *Journal of Counseling Psychology, 25*, 441 – 448.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective? *The Journal of Higher Education, 64*, 574 – 593.

Brown, P., Levinson, S. C. (1987). Politeness: Some universals in language use. Cambridge University Press, New York.

Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychology Review, 62*, 193 – 217.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245 – 281.

Cahasseigne, G., Grau, S., Mullet, E., & Cama, V. (1999). How well do elderly people cope with uncertainty in a learning? *Acta Psychologica, 103*, 229 – 238.

Chasseigne, G., Mullet, E., & Stewart, T. R. (1997). Aging and multiple cue probability learning: The case of inverse relationships. *Acta Psychologica, 97*, 235 – 252.

Choi, N., & Fuqua, D. R. (2003). The structure of the Bem sex role inventory: A summary report of 23 validation studies. *Educational and Psychological Measurement, 63,* 872-887.

Corbalan, G., Paas, F., & Cuypers, H. (2010). Computer-based feedback in linear algebra: Effects on transfer performance and motivation. *Computers & Education, 55,* 692 – 703.

Deffenbacher, K. A., & Hamm, N. H. (1972). An application of Brunswik's lens model to developmental changes in probability learning. *Developmental Psychology, 6,* 508 – 519.

Dykema, J., Price, J., DiLoreto, K., White, E., & Schaeffer, N. C. (2012). ACASI gender-of-interviewer voice effects on reports to questions about sensitive behaviors among young adults. *Public Opinion Quarterly, 76,* 311 – 325.

Evans, R. E., & Kortum, P. (2010). The impact of voice charateristics on user response in an interactive voice response system, *Interacting with Computers, 22,* 606 -614.

Fagot, B. I. (1985). Beyond the reinforcement principle: Another step toward understanding sex roles. *Developmental Psychology*, ***21***, 1097–1104.

Frey, M. C., & Detterman, D. K. (2004). Scholastic Assessment or *g*? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15,* 373 – 378.

Ginns, P. (2005). Meta-analysis of the modality effect, *Learning and Instruction, 15,* 313 -331.

Griggs, B. (2011, October 21). Why computer voices are mostly female, CNN. Retrieved from http://cnn.com/2011/10/21/tech/innovation/female-computer-voices/

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81-112.

Hawley, W. D., Rosenholtz, S. J., Goodstein, H. J. & Hasselbring, T. (1984). Good schools: what research says about improving student achievement. *Peabody Journal of Education 61*, 1-178.

Kanter, J., & Lindsay, S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389 – 406.

Karelaia, N., & Hogarth, R. M. (2008) Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404 – 426.

Kirschner, P. A., Sweller, J., & Ckar, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery,

problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75-86.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Kulhavy, R. W., & Stock, W. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279–308.

Kuncel, N. R., Credé, M., & Thomas, L. L. 2005. The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*, 63 -82.

Lysakowski, R. S., & Walberg, H. J. (1982). Instructional effects of cues, participation, and corrective feedback: a quantitative synthesis. *American Educational Research Journal, 19,* 559-578.

Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist, 40,* 25 – 265.

Mayer, R. E., & Anderson, R. B. (1992). The Instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology, 84,* 444 – 452.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in Multimedia Learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*, 312 – 320.

Mayer, R. E., Johnson, W. L., Shaw, E., & Sandhu, S. (2006). Constructing computer-based tutors that are socially sensitive: Politeness in educational software. *International Journal of Human-Computer Studies, 64*, 36 – 42.

Mclaren, B. M., Deleeuw, K. E., & Mayer, R. E. (2011). A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies, 69*, 70 -79.

Montazemi, A. R., & Gupta, K. M. (1998). On the effectiveness of cognitive feedback from an interface agent. *International Journal of Management Sciences, 25*, 643 – 658.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32,* 99-113.

Moreno, R. (2006). Does the modality principle hold for different media? A test of the methods-affects-learning hypothesis. *Journal of Computer Assisted Learning, 22*, 149 – 158.

Mayer, R. E., Moreno, R. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*, 358 – 368.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81 – 103.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B. & Dryer, D. C. (1995). Can Computer personalities be human personalities? International Journal of Human-Computer Studies, 43, 223-239.

Nass, C., Robles, E., Heenan, C., Bienstock, H., Trienen, M., 2003. Speech-based disclosure systems: effects of modality, gender of prompt, and gender of user. *International Journal of Speech Technology 6*, 113–121.

Nass, C., & Yen, C. (2010). *The man who lied to his laptop: What machines teach us about human relationships.* Oxford University Press, USA, 2010.

Pavio, A. (1986). *Mental representations: A dual coding approach.* New York: Oxford University Press.

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory and Cognition, 17*, 398 – 422.

Reed, S. K. (2006). Cognitive architectures for multimedia learning. Educational Psychologist, 41, 87 – 98.

Roberts, T. A. (1991) Gender and the Influence of Evaluations on Self-Assessments in Achievement Settings*, Pyschological Bulletin,* 109, 297-308.

Shneiderman, B., Plaisant, C., Cohen, M., & Jacobs, S. (2009). *Designing the user interface: Strategies for effective human-computer interaction.*Addison-Wesley Publishing Company, 2009.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153-189.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language, 19,* 129 – 146.

Tabbers, H. K., Martens, R. L., & van Merriënboer, J. J. G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology, 74,* 71 – 81.

Thomas, S., & Lane, D. (2010). *Gender differences in the effect of politeness in a computerized learning task*. Paper presented at the Southwestern Psychological Association conference, Dallas, Tx.

Todd, F. J., & Hammond, K. R. (1965). Differential feedback in two multiple-cue probability learning tasks. *Behavioral Science, 10*, 429–435.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies,* 66, 98-112.

Wiggins, J. S., & Pincus, A. L. (1989). The interpersonal circle: a structural model for integrating personality research. In R. Hogan, Ed. *Perspectives in Personality, 1,* 1-4.