Exploring User Satisfaction in a Tutorial Dialogue System

Myroslava O. Dzikovska, Johanna D. Moore
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom
m.dzikovska,j.moore@ed.ac.uk

Natalie Steinhauser, Gwendolyn Campbell
Naval Air Warfare Center Training Systems Division
Orlando, Florida, USA
gwendolyn.campbell,natalie.steinhauser@navy.mil

Abstract

User satisfaction is a common evaluation metric in task-oriented dialogue systems, whereas tutorial dialogue systems are often evaluated in terms of student learning gain. However, user satisfaction is also important for such systems, since it may predict technology acceptance. We present a detailed satisfaction questionnaire used in evaluating the BEETLE II system (REVU-NL), and explore the underlying components of user satisfaction using factor analysis. We demonstrate interesting patterns of interaction between interpretation quality, satisfaction and the dialogue policy, highlighting the importance of more fine-grained evaluation of user satisfaction.

1 Introduction

User satisfaction is one of the primary evaluation measures for task-oriented spoken dialogue systems (SDS): the goal of an SDS is to accomplish the task, and to keep the user satisfied, so that they will want to continue using the system. Typically, the PARADISE methodology (Walker et al., 2000) is used to establish a performance function which relates user satisfaction measured through questionnaires to interaction parameters that can be derived from system logs. This function can then be used to better understand which properties of the interaction have the most impact on the users, and to compare different system versions.

In contrast, tutorial dialogue systems are typically evaluated in terms of student learning gain, by comparing student scores on standardized tests before and after interacting with the system. This is clearly an important evaluation metric, since it directly assesses the benefit students obtain from using the system. However, it is also important to evaluate user satisfaction, since it can influence students' willingness to use computer tutors in a long run. Thus, recent studies have looked at factors that could influence user satisfaction in tutorial dialogue, such as different tutoring policies (Forbes-Riley and Litman, 2011), quality of speech output (Forbes-Riley et al., 2006), and students' prior attitudes towards technology (Jackson et al., 2009).

Assessing user satisfaction, however, is not a straightforward task. As we discuss in more detail in Section 2, user satisfaction is known to be a complex multi-dimensional construct, composed of largely independent factors such as perceived ease of use and perceived usefulness. Therefore, questionnaires used for assessing satisfaction need to be validated through user studies, and different satisfaction dimensions should be assessed independently. Therefore, SDS researchers are now starting to use techniques from psychometrics for this purpose (Hone and Graham, 2000; Möller et al., 2007). However, user satisfaction studies tutorial dialogue currently rely on simple questionnaires adapted from either task-oriented SDS or non-dialogue intelligent tutoring systems (Michael et al., 2003; Forbes-Riley et al., 2006; Forbes-Riley and Litman, 2011; Jackson et al., 2009), and these questionnaires have not been validated for tutorial dialogue systems.

In this paper, we make the first step towards developing a better user satisfaction questionnaire for tutorial dialogue systems. We present a user satis-

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 162–172,
Portland, Oregon, June 17-18, 2011.

faction evaluation of the BEETLE II tutorial dialogue system. Starting with a detailed user satisfaction questionnaire, we employ exploratory factor analysis to discover a set of dimensions for the students' satisfaction with a dialogue-based tutor. We then use the factors we derived to compare user satisfaction between two versions of our computer tutor that use different policies for generating the tutor's feedback. We investigate the relationships between the subjective satisfaction dimensions and the objective learning gain metric for the two systems. Finally, we carry out a more detailed investigation of our prior results on the relationship between user satisfaction and interpretation quality in tutorial dialogue. Our analysis also provides insights for further improving the questionnaire we developed and gives an example of how user satisfaction metrics developed for task-oriented dialogue can be adapted to different dialogue applications. It also opens new questions about how different properties of the interaction affect user satisfaction in tutorial dialogue, which can be investigated in future work.

The rest of the paper is organized as follows. We discuss the approaches for assessing user satisfaction with SDS in Section 2. In Section 3 we describe the BEETLE II tutorial dialogue system used in this evaluation. We describe our questionnaire design in Section 4, and describe its use in BEETLE II evaluation in Section 5. We conclude by discussing the implication of our analysis for tutorial dialogue system evaluation in Section 6.

2 Background

A typical approach to assessing user satisfaction in dialogue systems is collecting user survey data by asking users to rate their agreement with statements such as "the system was easy to use". In the simplest case of early PARADISE studies, the questionnaires contained 5 items assessing different dimensions of satisfaction, which were then summed to produce a total satisfaction score.

However, using simple questionnaires has drawbacks now recognized by the SDS community. First, if individual questions are expected to assess different dimensions of user satisfaction, they need to be validated first, or else they may be ambiguous and mean different things to different users. Second, summing or averaging over questions measuring different satisfaction components may not be the best approach, since it may conflate unrelated judgments (Hone and Graham, 2000).

To address this problem, SDS researchers have started using more complex questionnaires, where each underlying dimension of user satisfaction is assessed through multiple questions. Factor analysis is then used to determine which questions are related to one another (and therefore are likely to be assessing the same underlying satisfaction dimension), and to discard possibly ambiguous questions. Then, the PARADISE methodology can be used to relate different interaction parameters to individual components of user satisfaction.

Several such studies have been conducted recently (Hone and Graham, 2000; Larsen, 2003; Möller et al., 2007; Wolters et al., 2009), covering command-and-control and information-seeking dialogue. The questionnaires in those studies contained 25 to 50 items, and factor analyses typically resulted in 6- or 7-factor solutions, with dimensions such as acceptability, affect, system response accuracy and cognitive demand. The underlying factors found by those analyses tend to match up well, but not to overlap perfectly. In comparison, all user satisfaction questionnaires for tutorial dialogue systems that we are aware of contain 10-15 items which are either summed up for PARADISE studies, or compared individually to track system improvement (Michael et al., 2003; Forbes-Riley et al., 2006; Forbes-Riley and Litman, 2011; Jackson et al., 2009).

In this paper, we apply the more sophisticated SDS evaluation methodology to the BEETLE II tutorial dialogue system. We devise a more sophisticated user satisfaction questionnaire using SDS questionnaires for guidance and then apply factor analysis to investigate the underlying dimensions. We compare our results to analyses from two previous studies: SASSI (Hone and Graham, 2000), which is a validated questionnaire intended for use with a variety of task-oriented dialogue systems, and a more recent "modified SASSI" questionnaire which is a version of SASSI adapted for use with the INSPIRE home control system (Möller et al., 2007). Henceforth we will refer to this as INSPIRE.

3 BEETLE II Tutorial Dialogue System

The goal of BEETLE II (Dzikovska et al., 2010c) is to teach students conceptual knowledge in the domain of basic electricity and electronics. The system is built on the premise that encouraging students to explain their answers and to talk about the domain will lead to improved learning, a finding consistent with analyses of human-human tutoring in several domains (Purandare and Litman, 2008; Litman et al., 2009). BEETLE II has been engineered to test this hypothesis by eliciting contentful talk through explanation questions.

The BEETLE II learning material consists of two self-contained lessons suitable for college-level students with no prior knowledge of basic electricity and electronics. The lessons take 4 to 5 hours to complete, and consist of reading materials and interactive exercises. During the exercises, the students interact with a circuit simulator, building electrical circuits containing bulbs, batteries and switches, and using a multimeter to measure voltage. Then the tutor asks students to explain circuit behavior, for example, "Why was bulb A on when switch Y was open and switch Z was closed?" In addition, at different points in the lesson the tutor asks "summary" questions, asking students to define concepts such as voltage, and verbalize general patterns such as "What are the conditions that are required for a bulb to light?". At present, students use a typed chat interface to communicate with the system.[1]

We built and evaluated two versions of the system (Dzikovska et al., 2010a). The baseline non-adaptive tutor (BASE) requires students to produce answers, but does not provide any remediation and immediately states the correct answer. The fully adaptive version (FULL) engages in dialogue with the student, and tailors its feedback to the student's answer by confirming its correct parts and giving hints in order to help students fix missing or incorrect parts. The FULL system generates feedback automatically based on a detailed analysis of the student's input, and is capable of giving hints at different levels of specificity depending on the student's previous performance.

[1] A speech interface is being developed, but typed communication is common in online and distance learning, and therefore is an acceptable choice for tutorial dialogue as well.

These two system versions were designed to evaluate the impact of adaptive feedback (within the limitations of current language interpretation technology) on student learning and satisfaction. Our initial data analysis focused on the differences in student language depending on the condition (Dzikovska et al., 2010a), and on the impact of different types of interpretation errors on learning gain and user satisfaction (Dzikovska et al., 2010b). However, these initial results were based on an aggregate satisfaction score obtained by averaging over scores for all questions in our user satisfaction questionnaire. In this analysis, we take a more detailed look at the different factors that contribute to students satisfaction with the system, and their relationship with learning gain and interpretation quality.

4 Data Collection

4.1 Questionnaire Design

To support user satisfaction evaluation we developed a satisfaction questionnaire, REVU-IT (Report on the Enjoyment, Value, and Usability of an Intelligent Tutor). It consists of 63 items which cover all aspects of interaction with the tutoring system: the clarity and usefulness of the reading material; the graphical user interface to the circuit simulator; interaction with the dialogue tutor; and the overall impression of the BEETLE II system as a whole. The reading material, graphical user interface and interaction with the tutor sections are complementary, because they cover separate parts of the BEETLE II interface. We expect that all of these three components contribute to the overall impression score. For purposes of this paper, we will focus on the part of the questionnaire that relates to the natural language interaction with the tutor (REVU-NL), and its relationship to the overall impression score (REVU-OVERALL).

The REVU-IT questionnaire was developed by experienced cognitive psychologists (two of the authors of this paper). The REVU-NL section consists of 35 items shown in Appendix A. Its design was guided by questionnaires used in previous research, including INSPIRE and a questionnaire used to evaluate the ITSPOKE tutorial dialogue system (Forbes-Riley et al., 2006). REVU-NL contains a number of items from these, but omits items that are

not relevant to the BEETLE II domain (e.g, "Domestic devices can be operated efficiently with the system" or "The tutor responded effectively after I was uncertain"), and adds extra questions related to tutoring (e.g., "Our dialogues quickly led to me having a deeper understanding of the material"), based on the authors' previous experience in human factors research. We also slightly rephrased all questions to refer to "the tutor" rather than "the system".

The REVU-OVERALL section of REVU-IT consists of 5 items assessing the student's satisfaction with their learning as a whole. The questions are: "Overall, I am satisfied with my experience learning about electricity from this system."; "Working in this learning environment was just like working one-on-one with a human tutor"; "I would have preferred to learn about electricity in a different way."; "I would use this system again in the future to continue to learn about electricity."; "I would like to be able to use a system like this to learn about other topics in the future.". We use the averaged score over these 5 items to represent the student's overall satisfaction with the learning environment, referring to it as "overall satisfaction".

Adding new questions to the REVU-NL questionnaire on top of already existing questions is the initial step in addressing the issues discussed in Section 2: validating the individual questions and discovering the underlying dimensions of user satisfaction. Having a large number of questions asking about the same aspects of the interaction will allow us to group related questions together into dimensions ("factors"), and also to discover ambiguous questions that will need to be improved in future studies. The detailed discussion of the technique and issues involved is presented in Hone and Graham (2000).

4.2 Participants

We used REVU-IT as part of a controlled experiment comparing the BASE and FULL versions of the system. We recruited 87 participants from a university in the Southern US, paid for participation. Participants had little knowledge of the domain. Each participant signed consent forms and completed a pre-test, then worked through both lessons (with breaks), and then completed a post-test and a REVU-IT questionnaire. Each session lasted 3.5 hours on average.

Out of 87 participants that completed the study, 13 had an inordinate amount of trouble with interface: they typed utterances that could not be interpreted by the tutor (defined as having more than 3 standard deviations in interpretation errors compared to the rest), did not follow tutor's instructions or experienced system crashes. In addition, two participants were learning gain outliers (again, more than 3 standard deviations from average). These participants were removed from the analysis. The questionnaires from the remaining 72 participants are used in our data analysis.

5 Analysis

5.1 Underlying satisfaction dimensions

Each item in the REVU-NL questionnaire used a 5-point Likert scale, from "completely disagree" (1) to "fully agree" (5). Most of the items were phrased so that the agreement with the statement meant a positive evaluation of the system. For a few items, however, the polarity was reversed (e.g., "The tutor was not helpful"). Those items were reverse-coded, with 1 meaning "fully agree" and 5 "completely disagree", to ensure that a lower score on all questions corresponds to a negative assessment.

Following Hone and Graham (2000), we used exploratory factor analysis to group questionnaire items into clusters representing different dimensions. One of the standard approaches in determining how many factors ("question clusters") to use is the *scree test* which checks the number of eigenvalues in the question covariance matrix which are greater than 1. These typically correspond to principal components which reflect the underlying questionnaire structure. The scree test showed 7 eigenvalues greater than 1, resulting in the 7-factor solution presented in Table 1.

The loadings in the table are the correlation coefficients between the individual question scores and the variables representing the factors. Most of the correlations are quite high, indicating that the questions are strongly correlated both among themselves and the underlying factor. However, the last two factors contain only non-loading questions according to the criteria in (Hone and Graham, 2000), i.e., questions for which the correlations are too weak to be

#	Question	Loading
1	t29: Knew what to say at each point	0.82
1	t22: Easy to interact with the tutor.	0.79
1	t9: Not sure what was expected.	0.73
1	t18: Knew what to say to the tutor.	0.70
1	t14: The tutor was too inflexible.	0.69
1	t19: Able to recover easily from errors	0.69
1	t24: Easy to learn to speak to tutor.	0.69
1	t16: Tutor didn't do what I wanted.	0.65
1	t3: Tutor understood me well.	0.65
1	t15: Working as easy as with a human.	0.64
1	t13: Had to concentrate when talking.	0.62
2	t31 Tutor was an efficient way to learn.	0.79
2	t32: Easy to learn from the tutor.	0.78
2	t34: Tutor was worthwhile	0.72
3	t28: Tutor was irritating.	0.76
3	t10: Tutor was fun.	0.74
3	t7: Enjoyed talking with tutor.	0.72
3	t30: Dialogues were boring.	0.66
4	t2: Tutor took too long to respond	0.84
4	t33: Tutor responded quickly	0.84
5	t26: Didn't always understand tutor	0.89
6	(t3: The tutor understood me well)	0.4
7	(t25: Comfortable talking with tutor)	0.59

Table 1: Factors derived from the REVU-NL questionnaire, with question loadings for the factor to which each question was assigned. Question text shortened due to space limitations, full text presented in the appendix. Non-loading questions in parentheses.

reliable. In addition, factors 4 and 5 had fewer than 3 questions. Since the number of subjects in our data set is small, such factors may not be reliable. Therefore, we focus our remaining analysis on the top 3 factors from the questionnaire, each of which contains 3 or more questions.

Twelve questions in REVU-NL were "cross-loading" according to criteria in Hone and Graham (2000), that is, their two top loadings differed by less than 0.2. This indicates questions that are likely to be ambiguous, since they are strongly correlated with two (theoretically independent) variables. Such questions should be refined and re-designed in future surveys. These were questions *t1*, *t4*, *t6*, *t11*, *t12*, *t17*, *t20*, *t21*, *t23*, *t25*, *t27*, *t35* from the appendix. We removed them from our solution, and discuss the implications for survey design in Section 6.

The first component in our analysis lines up well with the *Transparency* and *Cognitive load* factors from INSPIRE, and *Response accuracy*, *Cognitive demand* and *Habitability* from SASSI, though it was not split into individual factors as in those analyses. We will refer to this factor as *Transparency*. The second component contains questions specific to tutoring. However, it is similar to the *Acceptability* dimension from INSPIRE (the original SASSI questionnaire did not include similar questions), which asked users to rate statements such as "domestic devices can be operated efficiently with the system". Thus, we will refer to it as *Acceptability*. Finally, our third dimension lines up best with the *Affect* and *Annoyance* items from SASSI.[2] We will refer to it as *Affect*.

Although the correspondences between our factors and those derived from SASSI and INSPIRE are not perfect, the fact that similar underlying factors are derived from different user groups and systems indicates that they are likely to be measuring the same underlying constructs.

5.2 Comparing satisfaction in different systems

Recall that in this study we combined the data from two systems: FULL, where the system provided students with adaptive feedback and hints, and BASE, where the system simply acknowledged the student's answers and then provided a correct answer without engaging in dialogue. Table 2 separates out the average factor scores for these two conditions, where a factor score is computed by averaging over scores of all questions assigned to that factor.

When comparing learning gain and overall satisfaction between the two systems (which is the overall impression of the system behavior as a whole, including circuit simulation and lesson design), the difference is *not* statistically significant (learning gain $t(69) = -0.95, p = 0.35$, overall satisfaction $t(69) = -1.52, p = 0.13$). In contrast, on individual dimensions related to tutoring the scores for BASE is significantly higher than the score for FULL (*Transparency*, $t(69) = -7.19, p < 0.0001$; *Acceptability*: $t(69) = -3.24, p < 0.01$; *Affect*:

[2]The *acceptability* dimension from INSPIRE is split between our factors 2 and 3, but most of the questions correspond to our factor 2 questions.

	FULL	BASE
Transparency	2.15 (0.56)	3.36 (0.81)
Acceptability	3.11 (1.02)	3.80 (0.77)
Affect	2.43 (0.80)	2.86 (0.996)
Overall	3.39 (0.88)	3.70 (0.83)
Learning gain	0.61 (0.15)	0.65 (0.22)

Table 2: Average scores for different satisfaction dimensions in FULL and BASE (standard deviation in parentheses)

$t(69) = -1.97, p = 0.05$). Comparing the means, the biggest difference in student ratings shows on the *Transparency* scale, while the affective reaction for the two systems is more similar (though still rated higher for BASE).

It is somewhat unexpected to see that the students were equally satisfied overall with both systems but rated the tutor in BASE more highly than in FULL, since the tutor behavior was the only thing different between conditions. We are at present investigating the reasons for this result. One possibility is that when students did not get much feedback from the tutor (as in BASE), other factors became more important to overall satisfaction, such as course design and quality of user simulation.

5.3 Relationships between subjective and objective outcome measures

We investigated the correlations between learning gain and different user satisfaction factors for the two system versions. Results are presented in Table 3. As can be seen from the table, learning gain and user satisfaction are only significantly correlated in FULL, and only for the acceptability and overall satisfaction factors. None of the factors in the BASE system correlate with learning gain. This indicates that the student's affective reaction to the system is not necessarily linked directly to its objective benefits. We discuss these results further in Section 6

5.4 Impact of interpretation quality on user satisfaction

It is generally known in SDS research that measures of interpretation quality such as word error rate and concept accuracy are strongly correlated with user satisfaction (e.g., (Walker et al., 2000; Möller et al., 2007)). Our system uses typed input and produces complex logical representations (rather than simple slot-value pairs), thus, these measures cannot be computed directly. However, in an earlier study we showed that another measure of interpretation quality, namely, percentage of utterances that could not be interpreted by the system ("uninterpretable utterances") is negatively correlated with learning gain and user satisfaction (Dzikovska et al., 2010b).[3]

	FULL	BASE
Transparency	0.32 (0.07)	0.06 (0.69)
Acceptability	**0.38** (0.03)	0.23 (0.16)
Affect	0.29 (0.08)	-0.10 (0.53)
Overall	**0.38** (0.02)	0.18 (0.28)

Table 3: Correlations between satisfaction factors and learning gain for two dialogue policies. Significance level in parentheses. Bold indicates significance at $p < 0.05$ level.

That study revealed an unexpected pattern. Although the system recorded the number of utterances it could not interpret in both FULL and BASE, students in BASE were never informed of any interpretation problems. Nevertheless, the proportion of such uninterpretable utterances was still significantly negatively correlated with user satisfaction in BASE. After analyzing correlations between different types of errors and user satisfaction, we hypothesized that this can be explained by the lack of alignment between the system and the student, in particular when students used terminology different from that used by the system (Dzikovska et al., 2010b).

We can now analyze this relationship in more detail, looking at correlations between interpretation problems and different components of user satisfaction. The results are presented in Table 4.

As can be seen from the table, the proportion of uninterpretable answers is significantly correlated with *Acceptability* in FULL, but not in BASE. This is not surprising, indicating that students who were told that they were not understood perceived the system as less useful for them. More surprisingly, *Transparency*, which is related to perceived ease of

[3] In that study, we computed user satisfaction with the tutor by averaging over the entire 35 questions in our questionnaire as an initial approximation.

	FULL	BASE
Transparency	-0.28 (0.1)	-0.25 (0.10)
Acceptability	**-0.58** (< 0.001)	-0.29 (0.07)
Affect	**-0.35** (0.04)	**-0.34** (0.04)
Overall	**-0.38** (0.03)	-0.27 (0.11)
Learning gain	**-0.38** (0.03)	-0.09(0.60)

Table 4: Correlations between satisfaction factors and uninterpretable utterances for two different policies. Significance level in parentheses.

use for the system, was not correlated with uninterpretable utterances. Finally, the proportion of uninterpretable utterances is significantly correlated with *Affect* for both systems. Moreover, the unexpected negative correlation we observed in the earlier study between satisfaction with the tutor and interpretation problems in BASE can be primarily attributed to the negative correlation with the *Affect* score.

6 Discussion

In this study, we attempted to apply insights from studies of user satisfaction in spoken dialogue systems to a different type of dialogue application: tutorial dialogue. We were looking to develop a better user satisfaction questionnaire for evaluating tutorial dialogue systems, and to implement an evaluation methodology which takes into account different underlying dimensions of user satisfaction.

The three dimensions we obtained based on exploratory factor analysis of REVU-NL align well with the dimensions reported in the SDS literature, which provides some evidence of their validity. However, the results are preliminary because of the small number of participants involved, and need to be replicated with additional participants and different tutoring systems. Regardless, our analysis highlighted important issues in designing satisfaction surveys for different dialogue genres.

When choosing which questions to include in a satisfaction questionnaire for a new system type, SASSI is a very attractive starting point, because it was validated across multiple SDS in two genres (command and control and information seeking). This also means that SASSI items are phrased very generally and therefore easier to adapt. In contrast, INSPIRE contains a number of questions specific to the command and control domain, asking whether the user thinks the system is useful in achieving their goals (i.e., operating the domestic devices). SASSI includes only one similar item, "The system was useful". It was classed as *Affect*, most likely because there were no other similar items. However, we think that such questions represent an important separate dimension, namely the "perceived usefulness" factor known to predict technology acceptance (Adams et al., 1989). Therefore we included several items in REVU-NL with similar intent, asking whether users thought the system was beneficial to their goal (i.e., learning the material). These items were clustered into a separate dimension by factor analysis, indicating that they should be included in other satisfaction surveys.

Moreover, some of the questions that appeared genre-independent to us proved to be cross-loading in our analysis, which is an indicator of ambiguity. Apparently, some of the items from task-oriented dialogue questionnaires did not transfer well. For example, statements like "The system didn't always do what I expected" are unambiguous for task-oriented dialogue, where the user is supposed to be in control of the interaction, and therefore has clear expectations of what the system should do. In contrast, in tutorial dialogue the tutor has control over the learning material. Thus, it may be more ambiguous as to what, if anything, students are expecting from the interaction.

Overall, our experience shows that it may not be possible, or indeed useful, to create completely generic surveys. However, we believe that questionnaires can be phrased generally enough to apply to a range of systems with similar goals, and REVU-NL in particular is useful starting point for comparing dialogue-based tutoring systems. We believe that the 18 questions that we retained as unambiguous in our analysis provide adequate assessment of user satisfaction, and are grouped into factors consistent with results of previous research. However, the questionnaire could be further improved by revisiting the cross-loading items we rejected as ambiguous, and seeing if their wording could be improved. We are also intending to use REVU-IT in evaluating a spoken version of BEETLE II, thus providing additional validation data on a different version of the interface.

With respect to evaluation methodology, our results highlight the need to look at different satis-

faction dimensions separately. We used our factors to further investigate a pattern that we discovered in previous research, namely, that students who speak in a way that is difficult for the system to interpret tend to be less satisfied with the tutor, *even when they are not told of the interpretation problems*. Looking at correlations with individual dimensions shows that this relationship is primarily explained by the *Affect* dimension. Our working hypothesis is that the lack of alignment between incorrect student answers and the answers supplied by the system caused students to perceive the system as a less likeable or cooperative conversational partner.

We also observed that *Acceptability*, but no other dimensions, were correlated with learning gain in FULL. One possible explanation is that students who are learning more believe that the system is helping them reach their goals (our definition of *Acceptability*). The FULL condition provides students with more explicit feedback as to their learning; whereas in BASE students may have a less accurate estimate of how well they are doing, and hence no satisfaction dimensions are correlated with learning gain.

It is worth noting that an earlier study investigating the relationship between user satisfaction and learning in two different tutorial dialogue systems (Forbes-Riley and Litman, 2009) found little correlation between the answers to individual questions on their satisfaction questionnaire and learning gain. Only one correlation, with the question "The tutor helped me to concentrate", reached significance in only one of the 4 conditions they investigated. This adds further evidence that the relationship between learning gain and satisfaction is not straightforward. However, our results are difficult to compare since the questionnaires used are different, and Forbes-Riley and Litman (2009) are studying correlations with individual questions rather than grouping related questions together. Developing better validated questionnaires will make such results easier to compare and interpret, and we believe that REVU-NL makes a significant step in that direction.

7 Conclusion and Future Work

In this paper, we proposed an improved questionnaire (REVU-NL) for evaluating user satisfaction in tutorial dialogue systems, which is an important evaluation metric alongside learning gain. We used the methodology from SDS evaluations to investigate different dimensions of user satisfaction, and their relationship to learning gain and different interaction properties. Next, we are planning to use the PARADISE methodology to establish predictive models that relate satisfaction dimensions to measurable interaction properties, so that we can determine development priorities, and make it easier to compare different system versions. We are also planning to collect additional questionnaire data with a speech-enabled version of the system, and verify our analyses on this extended data set.

Acknowledgments

This work has been supported in part by US Office of Naval Research grants N000141010085 and N0001410WX20278. We would like to thank our sponsors from the Office of Naval Research, Dr. Susan Chipman and Dr. Ray Perez, and the Research Associates who worked on this project, Katherine Harrison, Leanne Taylor, Charles Scott, Simon Caine, Elaine Farrow and Charles Callaway for their contribution to this effort.

References

Dennis A. Adams, R. Ryan Nelson, and Peter A. Todd. 1989. Perceived usefulness, ease of use, and usage of information technology. *MIS Quarterly.*, 13:319–339.

Myroslava Dzikovska, Natalie B. Steinhauser, Johanna D. Moore, Gwendolyn E. Campbell, Katherine M. Harrison, and Leanne S. Taylor. 2010a. Content, social, and metacognitive statements: An empirical study comparing human-human and human-computer tutorial dialogue. In *Sustaining TEL: From Innovation to Learning and Practice - 5th European Conference on Technology Enhanced Learning (EC-TEL 2010)*, pages 93–108, Barcelona, Spain, October.

Myroslava O. Dzikovska, Johanna D. Moore, Natalie Steinhauser, and Gwendolyn Campbell. 2010b. The impact of interpretation problems on tutorial dialogue. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics(ACL-2010)*, Uppsala, Sweden, July.

Myroslava O. Dzikovska, Johanna D. Moore, Natalie Steinhauser, Gwendolyn Campbell, Elaine Farrow, and Charles B. Callaway. 2010c. Beetle II: a system for tutoring and computational linguistics experimentation. In *Proceedings of the 48th Annual Meeting of*

the Association for Computational Linguistics (ACL-2010) demo session, Uppsala, Sweden, July.

Katherine Forbes-Riley and Diane J. Litman. 2009. Adapting to student uncertainty improves tutoring dialogues. In *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009)*, pages 33–40, Brighton, UK, July.

Katherine Forbes-Riley and Diane J. Litman. 2011. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, 25(1):105–126.

Katherine Forbes-Riley, Diane J. Litman, Scott Silliman, and Joel R. Tetreault. 2006. Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference*, pages 509–514, Melbourne Beach, Florida, USA, May.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6(3&4):287–303.

G. Tanner Jackson, Arthur C. Graesser, and Danielle S. McNamara. 2009. What students expect may have more impact than what they know or feel. In *Proceedings 14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK.

Lars Bo Larsen. 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In *Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, pages 209 – 214, December.

Diane Litman, Johanna Moore, Myroslava Dzikovska, and Elaine Farrow. 2009. Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. In *Proceedings of 14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK, July.

Joel Michael, Allen Rovick, Michael Glass, Yujian Zhou, and Martha Evens. 2003. Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments*, 11:233–262(30).

Sebastian Möller, Paula Smeele, Heleen Boland, and Jan Krebber. 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech & Language*, 21(1):26 – 53.

Amruta Purandare and Diane Litman. 2008. Content-learning correlations in spoken tutoring dialogs at word, turn and discourse levels. In *Proceedings of the 21st International FLAIRS Conference*, Coconut Grove, Florida, May.

Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. 2000. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*, 6(3).

Maria Wolters, Kallirroi Georgila, Robert Logie, Sarah MacPherson, Johanna Moore, and Matt Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interacting with Computers*, 21(4):276–287.

A REVU-NL **Questions**

t1 I felt in control of my conversations with the tutor.
t2 It took the tutor too long to respond to my statements.
t3 I felt that the tutor understood me well.
t4 The tutor didn't always do what I expected.
t5 The information that the tutor provided to me was incomplete.
t6 It was easy for me to become confused during our dialogue.
t7 I enjoyed talking with the tutor.
t8 The tutor interfered with my understanding of the topics in electricity and circuits.
t9 I was not always sure what the tutor expected of me.
t10 Conversing with the tutor was fun.
t11 It was easy to understand the things that the tutor said.
t12 The dialogue between me and the tutor was very repetitive.
t13 I had to really concentrate when I was talking with the tutor.
t14 The tutor was too inflexible.
t15 Working through the lessons with the computer tutor was as easy as working through the lessons with a human tutor.
t16 The tutor didn't always do what I wanted.
t17 I felt confident when talking with the tutor.
t18 I always knew what to say to the tutor.
t19 I was able to recover easily from errors during our dialogues.
t20 Talking with the tutor was frustrating.
t21 The information provided by the tutor was clear.
t22 It was easy to interact with the tutor.
t23 The tutor's dialogue was clumsy and unnatural.
t24 It was easy to learn how to speak to the tutor in a way that the tutor understood.
t25 I felt comfortable talking with the tutor.
t26 I didn't always understand what the tutor meant.
t27 The tutor was not helpful.
t28 I found conversing with the tutor to be irritating.
t29 I knew what I could say or do at each point in the conversation with the tutor.
t30 I found our dialogues to be boring.
t31 Having the tutor help me with the material was an efficient way to learn.
t32 It was easy to learn from the tutor.
t33 The tutor responded quickly.
t34 Having the tutor was worthwhile
t35 Our dialogues quickly led to me having a deeper understanding of the material.

B REVU-OVERALL **questions**

o1 Overall, I am satisfied with my experience learning about electricity from this system.
o2 Working in this learning environment was just like working one-on-one with a human tutor.
o3 I would have preferred to learn about electricity in a different way.
o4 I would use this system again in the future to continue to learn about electricity.
o5 I would like to be able to use a system like this to learn about other topics in the future.

C REVU-IT questions related to GUI and reading material (mentioned but not analyzed in the paper)

sl1	It was easy to navigate through the slides.
sl2	It took a long time for each new slide to be displayed.
sl3	The material on the slides was easy to understand.
sl4	The material on the slides was poorly written.
sl5	I would have benefited from more instrucion on how to move through the slides.
sl6	The material on the slides was interesting.
sl7	The slide navigation buttons didn't always work the way I expected them to.
sl8	The slides were annoying.
sl9	The material on the slides was written at a level far beneath my abilities.
sl10	I would prefer reading a text book over reading these slides.
e1	I found it difficult to learn how to build circuits and take measurements in the workspace.
e2	Completing exercises in the workspace was fun.
e3	Before beginning the lesson, I received the right amount of instruction on how to build circuits in the workspace and take measurements.
e4	The exercises were well designed to illustrate the important lesson concepts.
e5	Sometimes I didn't understand what I was supposed to do for an exercise.
e6	The method for connecting components with wires was counter-intuitive.
e7	Having to build all those circuits was annoying.
e8	I always knew exactly what to build and/or measure in the workspace, and how to do it.
e9	Circuits loaded quickly.
e10	Even if I didn't predict the outcome correctly ahead of time, once I completed an exercise, I always understood the point.
e11	It was easy to use the meter.
e12	There were more exercises than necessary to cover the lesson topics.
e13	I would have learned more if I had been able to build circuits with actual light bulbs and batteries.