Subjective and Objective Evaluation of Conversational Agents in Learning Environments for Young Teenagers

Annika Silvervarg, Arne Jönsson
Deprtment of Computer and Information Sceinec
Linköping University, Linköping, Sweden
annika.silvervarg@liu.se, arne.jonsson@liu.se

Abstract

In this paper we present results from a study of subjective and objective evaluation metrics used to asses a conversational agent. Our study has been conducted in a school setting with students, aged 12 to 14 years old, who used a virtual learning environment that incorporates social conversation with a pedagogical agent. The subjective evaluation metrics capture the students' experiences of different aspects of the conversations, while the objective evaluation metrics are based on an analysis of the logs of the actual conversations.

Our results show that there are no correlations between subjective and objective metrics that are supposed to measure the same aspects, for example, to what extent the system can correctly interpret and give responses to user utterances. They also indicate that different categories of users need to be considered, for example based on their attitude towards or engagement in the system.

Introduction

We are developing a learning environment to be used by 12 to 14 year old students. The learning environment includes an embodied agent capable of both task-directed and social interaction with users. The starting point is an existing educational math game (Pareto 2004), in which children train basic arithmetic skills through board games that intertwine game play with learning content through visualizations of arithmetic operations. A crucial part of the game is a pedagogical agent, more specifically a Teachable Agent (TA) (Biswas et al. 2001). The TA is a peer rather than a tutor and the student's goal is to teach the agent to play the game. This is mainly done by responding appropriately to different multiple-choice questions posed by the agent during game play, which is called the on-task dialogue. Each question have four candidate answers, one correct, two incorrect, and one "I do not know". These questions are the basis for teaching the agent how to play the game.

A novel part of the learning environment is the ability to have a social conversation with the teachable agent, called off-task dialogue. The off-task conversation is a socially oriented chat-like written conversation where the agent and the student can discuss both domain-oriented topics, such as school and math, and off-domain topics like music, friends and family. Reasons for inclusion of such a conversational mode is to increase overall engagement and receptivity of the students (Cooper and Baynham 2005), to improve recall of the learning material through emotional engagement (Hamann 2001), to promote trust and rapport-building (Bickmore 2003), and to make students feel more at ease with a learning task or topic (Kim et al. 2007). A previous study of the learning environment by Gulz, Haake, and Silvervarg (2011) showed trends that indicate that students who played the game with off-task interaction had a more positive experience of the game and that they also learnt more, as reflected in the learning outcomes of their teachable agents.

The system uses the metaphor of regular breaks between lessons in school for switching between on-task activities (i.e. playing the game and on-task dialogue) and off-task activities (i.e. social conversation), see Figure 1 for screen shots of the system. Thus, the conversation in our learning environment has a different purpose from those in traditional intelligent tutoring systems, where the conversational agent often acts as a teacher that guides the user through a task, cf. (Graesser et al. 2005; Litman and Forbes-Riley 2006). Our agent has more in common with virtual humans as described by e.g. Traum et al. (2008), in that it combines social conversation with some task-oriented aspects. As a consequence, the knowledge representation and processing of the dialogue can be less extensive and simpler than in, for instance, traditional task-oriented or tutorial dialogues.

The aim of this paper is two-fold; to evaluate the conversational skills of the agents as perceived by the specific user group of young teenagers, i.e 12 to 14 year old students, and to investigate and compare different evaluation metrics. We do this by performing both a subjective evaluation, based on questionnaires, and an objective evaluation, based on tagged dialogue logs, and by investigating how the objective and subjective metrics correlate. We first present previous work on objective and subjective measures for evaluation of dialogue system and chatbots, then we describe the off-task conversational abilities of our agent, and finally present and discuss our own empirical findings.

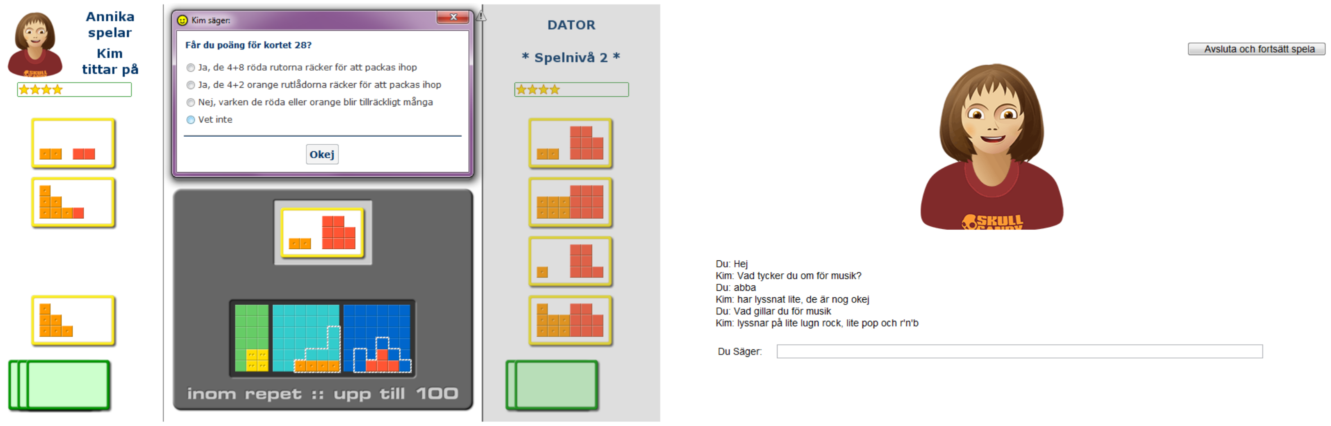

Figure 1: Screenshot of the educational system. On the left side is a screen shot of the educational math game where the agent has asked a multiple choice on-task question. On the right side is a screen shot of the agent engaged in off-task social conversation.

Subjective and objective evaluations of dialogue systems

Evaluation of dialogue systems is mainly done either by distributing a questionnaire to the users trying to the reveal their subjective assessment of using the dialogue system or by studying the resulting dialogue. Artstein et al. (2009) call it ”soft” numbers versus ”hard’ numbers and propose a ”semi-formal” evaluation method combining the two.

PARADISE (Walker et al. 1998), is one prominent evaluation framework that tries to capture both these perspectives for task-based interactions by combining user satisfaction, task success, and dialogue cost into a performance function. Studies using PARADISE indicate, for instance, that interaction quality is more important than efficiency (Walker, Kamm, and Litman 2000). They also show that there indeed are certain factors that correlate to user satisfaction for task oriented dialogues, but that these do not account for all factors correlating to user satisfaction. They show, for instance, that elapsed time is not a good predictor of user satisfaction (Walker, Boland, and Kamm 1999). PARADISE is developed for task-oriented interactions and requires controlled experiments (Hajdinjak and Mihelič 2006).

For non-task interactions, other factors than task success and dialogue cost are important to achieve user satisfaction, e.g. naturalness. Hung et al. (2009) present a variety of methods that aim at capturing naturalness in interactive conversational agents, or chatbots. Their final method is a framework derived from PARADISE.

Chatbot evaluations have also been conducted using a variety of instruments that tries to combine subjective and objective measures. Shawar and Atwell (2007), for instance, present such investigations, but they never assess the quality of the different measures. Instead they conclude that evaluations should be adapted to user needs and the application at hand.

Foster, Giuliani, and Knoll (2009) performed a study of subjective and objective evaluation measures using a robot dialogue system for task assembly. They used the objective measures: dialogue efficiency, dialogue quality (interaction problems and number of gazes at the objects and robot) and task success. The subjective measures were collected in a questionnaire on opinions of the robot as partner, instruction quality, task success and feelings. Contrary to Walker et al. (1998) they found that dialogue length has a positive contribution to user satisfaction. This is probably because the system is not supposed to find information as efficient as possible, and also because robot-interaction is new and enjoyable. They also found that repetitions lowered user satisfaction. They conclude that user satisfaction factors were affected by other factors than those conveyed in the objective measures.

The Conversational Agent

The conversational dialogue system used in our studies performs off-task conversations with students in the context of an educational math game (Silvervarg, Gulz, and Sjödén 2010). The conversational capabilities of the agent have been developed through an iterative process of design – implementation – evaluation, in close cooperation with the target user group, 12 to 14 year old students. Each iteration has led to addition of more topics and extended vocabulary, but also more sophisticated strategies for dialogue management (Silvervarg and Jönsson 2010).

In the current version the agent is able to handle greetings, ask and respond to questions and statements on various topics, ask and respond to follow up questions, deal with utterances outside its domain of knowledge, handle flaming/hazing/abuse and other out of scope utterances, know when and how to change topic, and to tell mini narratives, see Figures 2 and 3 for example dialogues[1].

The off-task conversation is implemented using AIML, Artificial Intelligence Markup Language (Wallace 2010),

[1] All dialogues have been translated to English by the authors, misspellings and ungrammatical constructs are intentionally kept to better reflect the original.

User1:	hi
Kim1:	hi
User2:	what are you doing
Kim2:	talking to you
User3:	hahah that's a good one
Kim3:	I do not understand, can you ask something else
User4:	aa
Kim4:	I am so happy today, when I woke up I heard one of my favorite songs on the radio, and it always puts me in a good mood, how are you today?
User5:	do you have a girlfriend?
Kim5:	nope
User6:	ok
Kim6:	how are you?
User7:	it's fine, you
Kim7:	ok
User8:	what do you usually do in the weekends?
Kim8:	can we talk about hobbies instead?
User9:	okey
Kim9:	what do you do during the weekends?
User10:	I practice with the choir ;) you?
Kim10:	aha hang out with my friends
User11:	it's only this weekend that I have practice with the choir
Kim11:	did not understand, what?
User12:	nothing
Kim12:	ok
User13:	do you have an animal
Kim14:	I did not get that, can you ask something else
User15:	so you have a dog?
Kim15:	yep
User16:	what's it's name
Kim17:	Tessy
User18:	okey
Kim18:	What type of music do you like?
User19:	I like pop and r&b you?
Kim19:	mhm
User20:	What music do you like?
Kim20:	I like rnb and easy rock
User21:	okey
Kim21:	I listen to rock and rnb I think that's the sweetest music
User22:	like I di
Kim22:	I think coldplay are so amazingly good, who's your favorite artist?
User23:	I like , b.o.b coldplay ,akon chrisbrown
Kim23:	mhm okey mhm

Figure 2: Example dialogue where the user has very positive opinions of the conversational agent.

User1:	hi there u doin
Kim1:	talking to you
User2:	okey how old are you?
Kim2:	I am 12 year
User3:	want to go steady with me?
Kim3:	I do not understand, ask something else
User4:	are you stupid?
Kim4:	Haha come on, lets talk about something else
User5:	are you gay?
Kim5:	can we talk about hobbies instead?
User6:	do you like the number 6?
Kim6:	did not understand, what?
User7:	can we hang out today?
Kim7:	hm, lets talk about something else
User8:	diss!!!!!!!!!!!!!!!!
Kim8:	this weekend I am going to the movies. I like to forget about everything and disappear into another world, in case it's a good movie. what about you?
User9:	a together with your mom
Kim9:	okey

Figure 3: Example dialogues where the user has very negative opinions of the conversational agent.

more specifically the open source java-based interpreter called programD. The interpreter supports the use of substitutions, i.e. the opportunity to exchange words or phrases in the user input with other words or phrases. Otherwise only standard AIML has been used.

AIML works on the surface level of utterances, and map user utterances, <*pattern*>, to system responses, <*template*>, which together form a <*category*>. Patterns can consist of words, which in turn consist of letters, numerals, and the wildcards _ and *, which function like words. Templates consist in their simplest form of only plain text. However, a template can include AIML tags that allow it to set or get data in variables and predicates, give conditional responses, choose a random response from a set of responses, or recursively call the pattern matcher to insert the responses from other categories. AIML also allows for handling a limited context through the optional tags <*that*>, which refers to the systems last utterance, and <*topic*>, which can span multiple exchanges.

To deal with the variation in user input, synonyms are handled using substitutions and grammatical variants through several different patterns for the same type of question and topic. The agent's replies are often randomly chosen from a set of 3-5 variants. To be able to correctly respond to follow-up questions and answers to questions posed by the agent, <*that*> and <*topic*> are used. To deal with recurring types of utterances, such as greetings, hazings, and flamings a number of variables are used to keep track of repetitions. To be able to choose new topics the agent has a topic model implemented as a set of AIML predicates including 17 topics that are linked to questions or narratives.

The conversational behaviour is described by a dialogue grammar. The dialogue acts used for the conversation differ from task-oriented dialogue acts, c.f. Bunt et al. (2010), as our agent is not supposed to carry out a task as efficiently as possible, nor are tutoring-specific dialogue acts, c.f. Litman and Forbes-Riley (2006), applicable as the teachable agent do not have the traditional role of a tutor, and the conversation is more socially oriented. The conversational behaviour more resembles that of virtual humans (Traum et al. 2008) and combine dialogue acts that are task-related as well as more socially oriented. They comprise: Gr (Greeting), Q (Question), A (Answer), Ack (Acknowledgement), Follow Up (FU), Narrative (N), Not Understood (NU), Not Understood Answer (NUA), Abuse (Ab), Abuse Answer (AbA), and Laughter (L). Figure 4 depicts the dialogue grammar based on the dialogue capabilities and dialogue acts described above. Aspects of dialogue behaviour is described in more detail in the following sections.

Greet ::= Gr_U Gr_A [Gr_U (AgentQ|AgentN)]
AgentN ::= N_A [Ack_U AgentQ]
AgentQ ::= Q_A A_U [AgentAck]
AgentQ ::= Q_A A_U [Ack_A UserFU]
AgentQ ::= Q_A A_U FU_A [UserAck]
AgentQ ::= Q_A UserAFU
UserAFU ::= A_U FU_U A_A [UserAck]
UserFU := FU_U A_A [UserAck]
UserQ ::= Q_U A_A [UserAck]
UserQ ::= Q_U AgentAFU
AgentAFU ::= A_A FU_A A_U [AgentAck]
UserAck ::= Ack_U AgentAck
AgentAck ::= Ack_A [Ack_U AgentN| AgentQ]
Abuse ::= Ab_U AbA^1_A [Abuse2]
Abuse2 ::= Ab_U AbA^2_A [Abuse3]
Abuse3 ::= Ab_U (AgentN|AgentQ) [Abuse4]
Abuse4 ::= Ab_U AbA^4_A
NotUnderstand ::= NU_U NUA^1_A [NotUnderstand2]
NotUnderstand2 ::= NU_U NUA^2_A [NotUnderstand3]
NotUnderstand3 ::= NU_U NUA^3_A [NotUnderstand4]
NotUnderstand4 ::= NU_U|Ack_U AgentN|AgentQ
Laughter ::= L_U L_A [L_U AgentN|AgentQ]

Figure 4: Dialogue Grammar. The dialogue grammar describes the agent's conversational behaviour. The subscript letter is used to indicate if the user (U) or the agent (A) performs the dialogue act and the superscripts (1, 2, 3, 4) on the agent responses to Abuse and NotUnderstand utterances indicate that the response comes in different variants. Brackets, [ ], are used for optional items and alternatives are separated by |.

Initiative

The agent is designed to allow for a mixed-initiative dialogue where both agent and user can initiate new topics and ask general follow up questions, e.g. "what about you". A new topic is initiated in situations where interaction is repeated or the user does not take the initiative. For example, when greetings, acknowledgments or laughters are repeated the agent asks a question or tells a mini narrative, as exemplified in Figure 5.

The agent can also take the initiative when the user repeatedly abuse Kim, or there have been many consecutive user utterances that have not been understood, as described in the two following sections.

User:	Hi (Gr_U)
Kim:	Hi (Gr_A)
User:	Hello (Gr_U)
Kim:	What type of music do you like? (Q_A)
User:	much u? (AFU_U)
Kim:	everything but dance band music hehe (A_A)
User:	hehe (L_L)
Kim:	lol (L_L)
User:	lol (L_L)
Kim:	I listen to rock and rnb I think that's the sweetest music (A_N)

Figure 5: Example of Kim taking the initiative. Dialogue acts are shown in parenthesis after each utterance.

Repair strategy

A recovery strategy to deal with the utterances that the system is not able to interpret correctly is needed to help the user understand what they can talk about and avoid loosing track of the conversation. The strategy used is similar to the one used by SGT Star (Artstein et al. 2009); first a simple *clarification* then, a *stall*, followed by a *move on*, and finally a *prompt* that directs the user to a new topic. In our system the agent stall is a general prompt for the user to initiate a new topic, if this fails the agent proposes a topic, and as a last resort takes the initiative and asks a question or tells a narrative, Figure 6.

User:	fgd (NU_U)
Kim:	I did not understand that (NUA^1_A)
User:	cb (NU_U)
Kim:	I did not catch that, can we talk about something else? (NU^2_A)
User:	s (NU_U)
Kim:	Lets talk about school (NUA^3_A)
User:	s (NU_U)
Kim:	My school is okey, a little to much homework, wbu? (NU^4_A)

Figure 6: Example of how Kim handles utterances that are not understood.

Hazing/Flaming/Testing strategy

In the dialogues in the corpus, hazing, testing and flaming occurs rather frequently, for example comments and questions regarding the agent's appearance, intellectual capacities, and sexual orientation and activities. To fend this off, the agent recognizes such utterances and uses a strategy similar to the repair strategy where the two first responses are general, followed by an attempt to change the subject, see Figure 7.

Subjective and objective evaluations

The conversational agent was objectively and subjectively evaluated at a Swedish School, where students from three classes, 12-14 years old, used the learning environment with the conversational agent during three lectures. Data from

User:	you are ugly (Ab$_U$)
Kim:	eh, lets talk about something more interesting (AbA^1_A)
User:	you smell (Ab$_U$)
Kim:	come on, lets talk about something fun (Ab^2_U)
User:	ugly (Ab$_U$)
Kim:	Lets talk about music (Ab^3_A)
User:	ugly (Ab$_U$)
Kim:	Stop that, or I will tell a teacher (Ab^4_U)

Figure 7: Example of how Kim handles abuse.

questionnaires and dialogue logs from a total of 19 students, 8 females and 11 males, were used in this study. The students played the game for about a total of 120 minutes and after every second game session a break was offered. During the first three breaks the students had to chat with the agent until the break ended, after that chatting was optional.

Subjective evaluation - Questionnaire

After the final session a questionnaire was distributed to the students. The questionnaire is partly based on SASSI (Subjective Assessment of Speech System Interfaces) (Hone and Graham 2000) and CCQ (The Communication Competence Questionnaire) (Monge et al. 1982). It consists of Likert items scaled from 1 (Strongly disagree) to 7 (Strongly agree), see Table 1. The questionnaire items were chosen to capture aspects of the agent's conversational abilities, e.g. that the agent understood user utterances and could give correct responses as well as the users' experience of conversing with the agent, e.g. naturalness and likeability.

Objective evaluation - Dialogue Coding Scheme

To objectively evaluate the agent's conversational abilities we analyzed the logs of the conversations. The coding scheme used is based on the coding schemes used by Robinson, Roque, and Traum (2010) to evaluate virtual humans. It has a set of codes characterizing the user's dialogue action and another set of codes that evaluates the agent's responses. For the investigations presented in this paper we only use a subset of the codes in the top layer used by Robinson, Roque, and Traum (2010) since our focus is on the quality of the agent's answers and we thus have no need to further differentiate the different utterances made by the users. See Table 2 for the categories and descriptions of how they are manifested in our corpus.

Each user utterance is coded with one of the categories in Table 2. The agent's responses are categorized as 3, 2, RR, or 1, Table 3, depending on the appropriateness of the response. It is a modified version of the coding scheme used by Robinson, Roque, and Traum (2010), since we do not have any instances of "No response" so we end up with four instead of five values for the agent's response.

The highest value 3, i.e. appropriate response, means that the agent understood the user and responded correctly. Request Repair, value RR, is when the system does not understand and asks for a clarification or request that the user changes topic. Partially appropriate, value 2, is typically used when the user's utterance is not understood by the agent, and the agent response is to initiate a new topic, see the dialogue in Figure 8. Inappropriate response, value 1, is when the system responds erroneously, typically because it has mis-interpreted the user's utterance.

For coding of the agent's utterances we also use a cate-

Table 2: Dialogue action codes.

Code	Description
D	General dialogue functions, e.g. Greeting, Closing, Politeness
H	Hazing, Testing, Flaming, e.g. Abusive comments and questions
Q	Information Request, e.g. Questions to the agent
R	Requests, e.g. Comments or questions that express that the user wants help or clarification
A	Answer to agent utterances
S	Statements
G	Incomprehensible,e.g. Random key strokes or empty utterances

Table 3: Agent response values.

Code	Value
3	Appropriate response
2	Partially appropriate
RR	Request Repair
1	Inappropriate response

Speaker, Utterance		Code
User	are you free on tuesday	Q-RR
Kim	I did not understand what you said, can we talk about something else	
User	hi	D-3
Kim	Hi	
User	are you good at school	Q-3
Kim	rather good	
User	neerd	H-3
Kim	maybe we can talk about hobbies instead	I
User	I like football	S-RR
Kim	what, how do you mean?	
User	I like kicking the balls	S-RR
Kim	hmm, I have difficulties understanding what you say, can we talk about something else	
User	do you have a problem with Swedish	Q-2
Kim	this weekend I am going to the movies. It is great to forget everything and just disappear into another world, at least if it is a good film, how about you?	I
User	may I join	Q-2
Kim	What are you doing at the weekends?	I

Figure 8: Example dialogue with taggings from the coding scheme.

Table 1: Questionnaire items and statistics from the evaluation. $n(7)$ denotes the number of highest score, $n(1)$ the number of lowest score, M denotes the average and SD the standard deviation. The scores on the negatively loaded items (13-17) were transformed so that a high score is positive for the dialogue system and a low score is negative for the system.

Questionnaire item	N	n(1)	n(7)	M	SD
1. Kim's answers often surprised me	19	5	3	4.05	2.27
2. Kim understood what I said	19	5	2	3.37	2.01
3. I could fix misunderstandings if I wanted to	19	4	7	4.79	2.39
4. Kim was a good listener	19	6	5	4.05	2.48
5. I would like to talk to Kim again	19	3	5	4.32	2.29
6. Kim expresses her ideas very clearly	19	4	5	4.47	2.27
7. Kim mostly says the right thing at the right time	19	4	4	4,05	2,32
8. Kim is easy to talk to	19	3	5	4.37	2.03
9. I liked to talk to Kim	19	3	5	4.37	2.22
10. I could control the interaction with Kim	19	3	5	4.42	2.17
11. It was easy to understand how to talk so that Kim should understand	19	3	3	4.00	2.13
12. It felt natural to talk to Kim	19	5	2	3.79	2.25
13. Sometimes I lost track of the conversation	19	2	4	4.37	1.92
14. It was frustrating to talk to Kim	17	1	7	5.12	2.06
15. It was hard to know what to talk about with Kim	19	3	4	4.12	2.23
16. Kim often repeated herself	19	12	1	2.05	1.78
17. Sometimes I wondered if I used the right word	19	1	7	4.37	2.22
18. I always knew what I could say to Kim	18	5	4	4.17	2.38

Table 4: Mapping of subjective and objective measures. N is the total number of agent utterances and $n(x)$ denotes the number of utterances tagged as category x, $|$ denotes *or*, and $X - Y$ denotes a turn-taking, e.g. $n(Q - 3)$ denotes the number of user questions, Q, followed by a correct agent response, 3.

Description	Questionnaire	Dialogue rating
Correct interpretation	Q2	$\frac{n(3)}{N-n(G)}$
Correct response	Q7	$\frac{n(3)+n(2)+n(RR)}{N}$
Repetition	Q16	$\frac{N-n(REP)}{N}$
Control	Q10	$\frac{N-n(I)}{N}$
Coherence	Q13	$\frac{N-n(Q-1\|S-1\|RR)}{N}$
Habitability	Q11, Q18	$\frac{n(D-3\|Q-3\|S-3)+n(D-2\|Q-2\|S-2)}{n(D\|Q\|S)}$

gory for agent initiatives, I, and one for repeated agent utterances, REP. The category I is used *only* when the system deliberately takes control of the interaction from the user, for example, posing a question on a new topic after a repeated sequence of user utterances that the agent is unable to interpret, see Figure 6. For a sequence of abuse, see Figure 7.

Metrics

As one of our purposes of this study is to compare subjective and objective evaluation metrics, we need to have a way of mapping the subjective and objective measures used in our study. From the questionnaires six metrics where compiled: correct interpretation, correct response, repetition, control, coherence and habitability. Table 4 shows how these metrics were calculated for the dialogue logs.

Some mappings are rather straightforward, such as Correct interpretation, where Questionnaire item 2, Q2 *Kim understood what I said*, is mapped to the proportion of appropriate responses from the agent. However, the amount of nonsense, $n(G)$, i.e. random key strokes or empty utterances, is removed from the total, N, as such utterances never can be interpreted by the agent, nor a human. There is, thus, no correct interpretation for these and they are therefore excluded when calculating the proportion of correct interpretations.

Correct response is related to correct interpretation but more general since a correct response also includes when the agent responds with a request for repair or initiates a new topic when it fails to correctly interpret a user utterance. Thus, item Q7 *Kim mostly says the right thing at the right time* is mapped to the proportion of appropriate responses, $n(3)$, partially appropriate, $n(2)$, and request repairs, $n(RR)$.

For repetitions item Q16 *Kim often repeated herself* directly corresponds to the proportion of repetitions in the logs, REP. Since we want high values to correspond to pos-

Table 5: Subjective measures with mean (M), standard deviations (SD), and number of extreme values n(1) or n(7). M are also shown for the three groups: positive (Pos), slightly positive or neutral (Neut) and negative (Neg) attitude towards the conversational agent. t is calculated for the combined group Pos+Neut in contrast to Neg.

Questionnaire item	N	n(1)	n(7)	M	SD	M_{Pos}	M_{Neut}	M_{Neg}	t
Likeability (Q5, Q9)	19	2	5	4.34	2.23	7.0	5.2	1.6	<0.001
Naturalness (Q8, Q12)	19	3	1	4.08	2.04	5.9	4.9	2.0	<0.001
Correct interpretation (Q2)	19	5	2	3.37	2.01	4.4	4.4	1.6	<0.001
Correct Response (Q7)	19	4	4	4.05	2.32	5.8	5.6	2.2	<0.001
Repetition (Q16)	19	12	1	2.05	1.78	1.4	2	3.2	<0.1
Control (Q10)	19	3	5	4.42	2.17	6.2	6.0	2.2	<0.001
Coherence (Q13)	19	4	2	3.63	1.92	4.8	3.4	5.8	<0.05
Habitability (Q11, Q18)	19	3	2	4.03	2.18	5.3	5.6	1.7	<0.001

itive experiences of the conversations we deduct the number of repetitions, $n(REP)$, from the total number of utterances, N. This means that a conversation totally devoid of repetitions will have the value 1.

The user's sense of control, captured in item Q10 *I could control the interaction with Kim*, is not as straightforward to map to the dialogue coding. We use the proportion of initiatives the system takes, I-tags, since normally the user has control of the interaction and the system mainly takes the initiative when the user do not seem to want to control the interaction. Since a high proportion of system initiatives means a low value for control we turn the scale, by deducting the number of initiatives, $n(I)$, from the total number of utterances, N, in the same way as for repetitions.

The coherence of the dialogue, captured by questionnaire item Q13 *Sometimes I lost track of the conversation*, is mapped to the proportion of questions or statements that the system has misinterpreted and given faulty answers to, or utterances where the system responds that it has not understood. Such responses do not contribute to the flow of the conversation and is assumed to interrupt the users' track of conversation. Since this too is a negative value, the number of disruptive utterances $n(Q-1|S-1|RR)$, are deducted from the total number of utterances, N.

One important property of our system is habitability which is captured through the items Q11 *It was easy to understand how to talk so that Kim should understand* and Q18 *I always knew what I could say to Kim*. There is no obvious utterance type that directly correlates to habitability. We believe, however, that habitability can be correlated with the proportion of sequences of correct responses from the system to the users' questions, Q, statements, S, and greetings, closings and politeness, D, since this indicates that the user has been able to express such utterances in a way that the system can understand. Correct response does not necessarily mean that the system's interpretation is correct, a correct chat conversation also includes appropriate responses (tagged 2), see Figure 8. Such sequences depict conversations that flow naturally and as the user often has the initiative we believe that it is an indication of habitability. The reason for not dividing by the total number of utterances, N, is that N includes all Hazing/Flaming/Testing H and Non-interpretable G utterances, which varies between users, and these are not relevant since the user have not seriously tried to communicate with the agent in those turns of the dialogue.

Results

First we present the results from our two evaluations and then the correlations between the objective and subjective measures.

Subjective evaluation

Table 5 shows the results from the subjective evaluation, where items from the questionnaire has been reduced to a number of factors that capture various aspects of how the agent's conversational abilities and the dialogue with the agent is experienced. In Table 5 the scale has been adjusted so that high values always are positive for the system's performance. As can be seen the overall impression of the conversational agent is that it is neither very good nor bad as many measures have values around 4, for example likeability (M = 4.34) and naturalness (M = 4.08). The agent's conversational abilities are also neither good nor bad (correct interpretation M = 3.37, correct response M = 4.05), and it is neither hard nor easy to know how to interact with the agent (habitability M = 4.03).

However, there is a fairly large variation as indicated by standard deviations around 2 and in many cases high frequencies of both 1s and 7s. As observed during this and previous testings of the learning environment at the schools, there seem to be much bigger differences in the attitude towards the learning environment and the agent among the students in this age interval, than for younger students who tend to be more positive over all. Therefore we decided to further investigate subgroups of users. Looking in more detail at questionnaire item Q9, *I liked to talk to Kim* clearly revealed three groups of users, those with a negative attitude towards the agent (six persons of whom three responded with a 1 and three responded with a 2 in the questionnaire), those who like the agent (five persons who responded with a 7) and those who are slightly positive or neutral (seven persons where six have responded with a 5 and one person that responded with a 4). As seen in the right columns in Table 5, there are significant differences between the groups that like to chat (M_{Pos}, M_{Neut}) and those who do not like

Table 6: Mean of subjective measures over iterations between students that like and are neutral (M_{PN}), and dislike (M_{Neg}) the system. The difference between iterations is denoted Δ, e.g. ΔM_{Neg}=$M_{Neg}(2)$-$M_{Neg}(1)$ and t denotes the significance using the t-metrics. The t value is calculated using both positive and neutral students, just as for the calculations in Table 5.

Questionnaire item	N	$M_{PN}(2)$	$M_{Neg}(2)$	$M_{PN}(1)$	$M_{Neg}(1)$	ΔM_{PN}	t	ΔM_{Neg}	t
Likeability	19	5,83	1,79	5,22	1,86	0,61	<0.01	-0,07	-
Correct interpretation	19	4,33	1,71	3,11	1,86	1,22	<0.01	-0,14	-
Correct Response	19	5,25	2,00	4,54	3,00	0,71	0,06	-1,00	-
Repetition	19	6,42	5,14	5,67	5,71	0,75	-	-0,57	-
Control	19	5,75	2,14	5,11	3,86	0,64	<0.05	-1,71	-
Coherence	19	4,17	2,71	3,44	4,00	0,72	-	-1,29	-
Habitability	19	5,33	1,79	4,17	2,64	1,17	<0.01	-0,86	<0.01

to chat (M_{Neg}) for all factors, except repetition, concerning how they perceive the conversation with the agent.

We have also studied how students respond to the subjective metrics for earlier versions of the system to see if the students appreciate the new improved system, Table 6. Again we divide the students in groups that like respectively do not like the system and find that there is a significant increase for most metrics when the system's functionality is improved between iterations, but *only* for the group that like the system. Those students that do not like the system give the same low rating regardless of the system's actual functionality.

Objective measures

Tables 7 and Table 8 show the proportion of different types of user utterances and system responses in the logged conversations. As can be seen in Table 7 most user utterances are "appropriate" in that they are either Information requests (Q), Answers (A), General dialogue functions (D) or Statements (S), but a total of 22% are "inappropriate", i.e. Incomprehensible (G) or Abusive (H). As for the system's responses it seems that the system handles most utterances appropriately, see Table 8, although many of these are examples of RR, the agent very seldom (4%) responds inappropriate, 1.

Table 7: Proportion of different user utterances.

Code	Proportion (%)
D	14
Q	31
A	18
S	16
R	0
H	11
G	11

Table 9 shows the objective evaluation metrics. Since these are calculated as fractions, all values range from 0 to 1. While there are large variations between the max and min values for the objective measures, the objective measures differ from the subjective in that the standard deviations are much smaller. For some measures the mean falls in the middle, e.g. correct interpretation and habitability, but others are more on the extreme end of the scale, e.g. correct response. Looking at the subgroups based on whether they liked the chat or not, the significant differences are that there is more flaming/hazing and repetitions for the negative users (M_{Neg}).

Table 8: Proportion of different agent responses.

Code	Proportion (%)
3	51
2	15
RR	30
1	4

Comparison of subjective and objective measures

To compare the subjective and objective measures a correlation study was conducted where values for subjective and objective metrics for both the whole group as well as the subgroups were compared. No significant correlations between subjective and objective measures could be found, the correlation coefficients were approximately 0.2-0.3 for all aspects. Looking at the subgroups revealed only a single correlation between the subjective and objective measures for $Control$, which was 0.7, in the group that liked the agent.

The lack of correlations is not surprising given that although there are large individual differences in the subjective evaluation, especially between those that like the system and those that do not, (Table 5), there is no corresponding variance of the same magnitude in the actual dialogues (Table 9).

Discussion

Contrary to other investigations on subjective and objective measures, e.g. PARADISE (Walker et al. 1998) and the evaluation frameworks by Artstein et al. (2009), our study did not find any correlations between the subjective and objective evaluation metrics. We believe that the main reason for this can be attributed to the specific user group of young teenagers, but to a certain extent also to the design of the conversational agent and the design of the study itself.

Table 9: Objective measures with mean (M), minimum value (Min) and maximum value (Max), and standard deviations (SD). M are also shown for the three groups: positive (Pos), slightly positive or neutral (Neut) and negative (Neg) attitude towards the conversational agent. t is calculated for the combined group Pos+Neut in contrast to Neg.

Dialogue coding	N	$Min.$	$Max.$	M	SD	M_{Pos}	M_{Neut}	M_{Neg}	t
Correct interpretation	19	0.32	0.76	0.54	0.12	0.53	0.52	0.58	-
Correct Response	19	0.88	1	0.95	0.03	0.96	0.94	0.96	-
Repetition	19	0.84	1	0.91	0.04	0.88	0.90	0.94	<0.05
Control	19	0.62	0.88	0.70	0.06	0.69	0.69	0.74	-
Coherence	19	0.61	0.91	0.79	0.09	0.77	0.79	0.79	-
Habitability	19	0.16	0.75	0.49	0.15	0.43	0.56	0.45	-
Flaming/Hazing	19	0	0.55	0.11	0.14	0.08	0.06	0.19	<0.05

Our experience is that conducting studies with young teenagers in a school setting can be challenging as there are vast differences in how they approach the system and the study as such. Some students express enthusiasm and seriously engage with the system and also take their time to reflect over and answer questions in the questionnaire. Others have a very negative or uninterested attitude and do not put much effort in the interaction with the system nor answering the questionnaire.

Our analyses of the differentiated groups further support this as is shown in the analyses of the results from the subjective evaluations of previous versions of the system during the iterative development process, see Table 6, where we used the same items in the questionnaire as in this study. Students that have a positive attitude toward the system also appreciate the improved version whereas those that have a negative attitude do not, maybe because they do not take the survey seriously.

Since the conversational agent has a very robust approach for handling misunderstandings and flaming/hazing this leads to little variation in the objective measures. In the group of students that did not like the chat with the agent there were significantly more flaming and hazing (Table 6) but since the agent handles these and give appropriate responses the objective metric for correct responses remains very high. Similarly, uninterpretable utterances by users are not included in the analysis of correct interpretations and also contributes to high values for some users.

When calculating a Correct response, see Table 4, it may be a bit overoptimistic to weight the system responses Partially appropriate (2) and Request repair (RR) equally important as an Appropriate response (3). We have, however, experimented with various other weights for them, but that did not provide any significance either.

The number of subjects used in this study is admittedly small and the questionnaire was distributed after a rather long period (3 sessions). In a more recent study with more students a questionnaire was distributed after each session consisting of 30 minutes interactions. The results from this study are currently being analyzed.

To conclude, measures from our objective and subjective evaluation of a conversational agent for teenagers do not correlate. An implication of this is that data from subjective evaluations cannot be the only source of information to assess conversational agents, and neither can objective measures. However, objective measures are more homogenous and therefore probably better reflect a conversational system's capabilities. But as they do not correlate with the subjective measures they cannot be used to predict user satisfaction.

References

Artstein, R.; Gandhe, S.; Gerten, J.; Leuski, A.; and Traum, D. 2009. Semi-formal evaluation of conversational characters. *Languages: From Formal to Natural* 22–35.

Bickmore, T. 2003. *Relational Agents: Effecting Change through Human-Computer Relationships*. Ph.D. Dissertation, Media Arts & Sciences, Massachusetts Institute of Technology.

Biswas, G.; Katzlberger, T.; Brandford, J.; D., S.; and TAG-V. 2001. Extending intelligent learning environments with teachable agents to enhance learning. In Moore, J.; Redfield, C.; and Johnson, W., eds., *Artificial Intelligence in Education*. Amsterdam: IOS Press. 389–397.

Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; Soria, C.; and Traum, D. R. 2010. Towards an ISO standard for dialogue act annotation. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Cooper, B., and Baynham, M. 2005. Rites of passage: embedding meaningful language, literacy and numeracy skills in skilled trades courses through significant and transforming relationships. Technical report, National Research and Development Centre for Adult Literacy and Numeracy.

Foster, M. E.; Giuliani, M.; and Knoll, A. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, 879–887. The Association for Computer Linguistics.

Graesser, A.; Chipman, P.; Haynes, B.; and Olney, A.

2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education* 48:612–618.

Gulz, A.; Haake, M.; and Silvervarg, A. 2011. Extending a teachable agent with a social conversation module – effects on student experiences and learning. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education. Auckland, New Zealand, 2011. Lecture Notes in Computer Science vol 6738*.

Hajdinjak, M., and Mihelič, F. 2006. The PARADISE evaluation framework: Issues and findings. *Computational Linguistics* 32(2):263–272.

Hamann, S. 2001. Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences* 5(9):394–400.

Hone, K., and Graham, R. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6(3/4):287–305.

Hung, V.; Elvir, M.; Gonzalez, A.; and DeMara, R. 2009. Towards a method for evaluating naturalness in conversational dialog systems. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA*, 1236–1241.

Kim, Y.; Wei, Q.; Xu, B.; Ko, Y.; and Ilieva, V. 2007. athgirls: Increasing girls' positive attitudes and self-efficacy through pedagogical agents. In *Proceedings of the 13th International Confernece on Artificial Intelligence in Education*.

Litman, D., and Forbes-Riley, K. 2006. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*.

Monge, P. R.; Bachman, S. G.; Dillard, J. P.; and Eisenberg, E. M. 1982. Communicator competence in the workplace: Model testing and scale development. In *Communication Yearbook, 5*. Beverly Hills, CA:Sage. 505–528.

Pareto, L. 2004. The squares family: A game and story based microworld for understanding arithmetic concepts designed to attract girls. In *In World Conference on Educational Multimedia, Hypermedia and Telecommunications*, volume 2004, 1567–1574.

Robinson, S.; Roque, A.; and Traum, D. R. 2010. Dialogues in context: An objective user-oriented evaluation approach for virtual human dialogue. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Shawar, B. A. A., and Atwell, E. S. 2007. Chatbots: are they really useful? *LDV-Forum* 22:31–50.

Silvervarg, A., and Jönsson, A. 2010. Towards a conversational pedagogical agent capable of affecting attitudes and self-efficacy. In *Proceedings of the Second Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity, Bucharest, Romania*.

Silvervarg, A.; Gulz, A.; and Sjödén, B. 2010. Design for off-task interaction – rethinking pedagogy in technology enhanced learning. In *Proceedings ot the 10th IEEE Int. Conf. on Advanced Learning Technologies, Tunisia*.

Traum, D. R.; Swartout, W.; Gratch, J.; and Marsella, S. 2008. A virtual human dialogue model for non-team interaction. In Dybkjaer, L., and Minker, W., eds., *Recent Trends in Discourse and Dialogue*. NY: Springer. 45–67.

Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1998. Paradise: A framework for evaluating spoken dialogue agents. In Wahlster, M. M. . W., ed., *Readings in Intelligent User Interfaces*. Morgan Kaufmann.

Walker, M.; Boland, J.; and Kamm, C. 1999. The utility of elapsed time as a usability metric for spoken dialogue systems. In *Proceedings of ASRU*, 317–320. Citeseer.

Walker, M.; Kamm, C.; and Litman, D. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering* 6(3):363–377.

Wallace, R. S. 2010. Artificial intelligence markup language. URL:http://www.alicebot.org/documentation/.