Research Article

Student Evaluations of a (Rude) Spoken Dialogue System Insights from an Experimental Study

Regina Jucks , Gesa A. Linnemann , and Benjamin Brummernhenrich

University of Muenster, Institute of Psychology for Education, Germany

Correspondence should be addressed to Regina Jucks; jucks@uni-muenster.de

Received 21 January 2018; Revised 1 June 2018; Accepted 11 June 2018; Published 1 August 2018

Academic Editor: Thomas Mandl

Copyright © 2018 Regina Jucks et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Communicating with spoken dialogue systems (SDS) such as Apple's Siri® and Google's Now is becoming more and more common. We report a study that manipulates an SDS's word use with regard to politeness. In an experiment, 58 young adults evaluated the spoken messages of our self-developed SDS as it replied to typical questions posed by university freshmen. The answers were either formulated politely or rudely. Dependent measures were both holistic measures of how students perceived the SDS as well as detailed evaluations of each single answer. Results show that participants not only evaluated the content of rude answers as being less appropriate and less pleasant than the polite answers, but also evaluated the rude system as less accurate. Lack of politeness also impacted aspects of the perceived trustworthiness of the SDS. We conclude that users of SDS expect such systems to be polite, and we then discuss some practical implications for designing SDS.

1. Introduction

Advances in speech recognition move our interactions with spoken dialogue systems (SDS), such as Apple's Siri or Google Now, ever closer to human dialogue. Over and above the capability of conversing in natural language, one strain of development concerns genuinely social aspects of human communication, such as alignment, addressing by name, and politeness. This paper addresses how students evaluate the communication behavior of an SDS that employs the social strategies of politeness and rudeness. Their evaluations address variables such as acceptance, appropriateness, and competence, all of which are also relevant in evaluating human behavior. In the following, we briefly address the technological aspects of SDS. Then we outline politeness theory and introduce how speakers take their speech partners' autonomy and affiliation into account when formulating messages. Finally, we provide results of our empirical research on the perception of SDS, which offer insights about how to design them effectively.

1.1. Technology That Interacts: Insights into the Mechanisms and Usage of Spoken Dialogue Systems. More and more

computers are now able to communicate with their users in natural language [1]. For example, spoken dialogue systems (SDS) serve as personal assistants and are implemented in smartphones (like Siri from Apple and Cortana® from Microsoft), cars [2], or special devices (such as Echo from Amazon: [3]). SDS are also used in educational contexts, e.g., in learning programs for children [4].

The first attempts at emulating human communication were simple chatbots such as Eliza which tried to emulate a psychotherapist in the Rogerian, person-centered tradition [5]. How chatbots can be made more human-like is still a very active field of development, for example, by evaluating the conditions of *uncanny valley* reactions in animated and text-based bots (e.g., [6]). Since 1990, the Loebner Prize has been held as an annual competition with an award given to the creator of the chatbot that most convincingly acts as a human interlocutor.

Whereas chatbots are usually text-based and often have entertainment purposes, SDS serve as an interface to a specific system and let computer systems and human users interact using spoken language; that is, the computer systems are capable of understanding and producing spoken language. This can reach from rather simple "command-and-control"

interactions, which refer to short, single controlling functions (e.g., [4]), to systems that are able to handle complex input and to generate complex, natural language (e.g., [7, 8]).

SDS have become much better at mimicking human interactions. Incrementality allows the system to implement back-channels and barge-ins. These are associated with facilitated *grounding* [9] in human-human interaction, meaning that a user is more confident about whether the system shares the user's understanding of the topic of the conversation and the meaning of the words used in it [7]. Connecting and operating SDS via the Internet can help support the system and the user in a concrete situation, e.g., by searching the Internet for the answer the SDS does not "know" directly. Additionally, such searches can be analyzed and used to improve the SDS [1]. Due to their proactive behavior, SDS are no longer restricted to simple replies. They can now initiate conversations themselves and provide information without being asked [1]. Thus, concerning their language, SDS can possess a high degree of anthropomorphism [10–12].

Providing these characteristics, SDS can be employed as tutors, termed intelligent tutoring systems. Graesser and colleagues [13] recently introduced an intelligent tutoring system from a new, "mixed-initiative" generation. Their conversational agent is able to maintain interactions over multiple turns, presenting problems and questions to the learner. Thus, it promotes active knowledge construction and outperforms mere information-delivery systems. Besides its conversational capabilities, social aspects of SDS play an important role. Research on intelligent tutoring systems has recently turned its attention on agents that can act socially and respond to tutees' affective states (e.g., [14, 15]). When intelligent tutors and tutees are teammates, for instance, tutoring seems to be more effective [16]. Put differently, how intelligent tutors—and SDS in general—communicate plays an important role. Politeness has seemingly been neglected in conversational agents [17]. However, some computer tutors already exhibit different kinds of polite instruction [18]. In the next section, we will give an overview of politeness theory and its effects on communication.

1.2. Mitigating Face Threats: Insights into Politeness Theory. Communication involves more than just conveying information. It is a social activity that contributes to individuals' social needs and impacts their self-concepts. In their politeness theory, Brown and Levinson [19] argue that every person has a public self-perception, the so-called *face*, and that this face needs to be shaped positively. Individuals' needs for belonging and support affect the *positive face*, and the need for autonomy and freedom of action affect the *negative face*. Communication behavior that serves as support of a person's face is termed *face work*.

Actions that harm and affect someone's face are called *face-threatening acts* (FTAs). To what extent an FTA is perceived as serious depends on social distance, power relations, and the absolute imposition. During communication, every single contribution might include FTAs, be it through direct orders, which restrict autonomy and independence, or through corrections, which in a way serve as a rejection of the person. Politeness theory describes several strategies

for how to mitigate an FTA [20]. The first option is not to perform the FTA and, accordingly, the utterance (the topic/aspect is simply not addressed). Second, the FTA can be transmitted *off record*; that is, the speaker remains vague or ambiguous. Hence he/she cannot be sure if the intended meaning is conveyed. Third, negative politeness is used, which can reduce the direct imposition on the hearer (e.g., through apologizing, employing hedges, and being conventionally indirect). This includes formulating a request indirectly as a suggestion. Fourth, positive politeness aims to meet the hearer's need for belonging and at gratification. A message might be started with positive feedback on the communication partners' intelligence. Mitigating face threats whenever autonomy or affiliation is threatened is a natural communication behavior.

1.3. Human-Like Interaction? Empirical Research on the Perception of SDS. As SDS possess enormous capabilities, users are likely to consider them conversational partners. SDS are perceived with their human-like qualities even when users are aware that they are communicating with a computer [21]. Basic language capabilities, for example, giving simple responses such as "yes" and "no", seem to be enough to perceive the computer as a human-like being [22]. When users perceive the SDS as a competent partner, they also perceive it as a social actor [23]. In this way, an SDS with a female voice can, for example, potentially activate corresponding gender stereotypes [10]. Thus, it becomes relevant how users assess an SDS according to different aspects that humans are evaluated on. One aspect that also directly influences how the information conveyed by the SDS is processed is *trustworthiness*, which includes *ability*, *benevolence*, and *integrity* (see [24] and the ABI model; [25]).

If computers are perceived as social agents, this influences our interaction with them. For example, people prefer systems that communicate in a personal manner [21]. People also usually communicate politely with computers and avoid explicit face threats (e.g., [26]) but also sometimes lie or behave intentionally rudely toward them [27]. This could be caused by disinhibition but could also be a reflection of the prevalence of rudeness in human communication [28].

Given that people tend to treat computers as humans, are social behaviors such as politeness also expected from SDS [29]? In human communication, politeness improves social perceptions. Polite speakers appear more likable and recipient oriented [30, 31]. Some studies have also shown effects on attributes such as perceived integrity or competence [32]. Thus, if an SDS is perceived similarly to a human interlocutor, similar effects should be found.

Regarding trust, it can be argued that SDS can be conceptualized as receivers of trust, as trustees (see also [33]). McKnight [34] has stated that "trust in technology is built the same way as trust in people" (p. 330). In the vast majority of conceptualizations, trust incorporates the willingness to be vulnerable [35] and therefore the willingness to depend on somebody else (e.g., [36]).

Regarding politeness, users tend to communicate politely with computer systems (e.g., [26]). Pickard, Burgoon, and Derrick [37] have found that likeability, in turn, increases the

tendency to align, e.g., using the same words and expressions to a conversational agent. Lexical alignment is perceived as polite [38, 39] and evokes positive feelings [38, 40]. Gupta, Swati, Walker, and Romano [41] developed a system which employs artificial spoken language and politeness principles in task-oriented dialogues. They found, for example, that the predictions of politeness theory are applicable to discourse with such systems: the strategies had an impact on the perception of politeness (see also [42]). However, we are unaware of any investigations into intentionally rude systems, with the exception of a rude intelligent tutoring system that has proven beneficial for some, but not all, of its students [43].

Nowadays, companies offer the possibility to personalize SDS (e.g., Siri can be taught the user's name and relationships; navigation systems' language style can be adjusted, e.g., restricted and elaborated language style: NIK-VWZ01 Navigation, n.d.). De Jong, Theune, and Hofs [44] developed an embodied conversational agent that is able to align to the politeness shown by its interlocutor. Some participants appreciated that the system aligned in terms of politeness, others preferred a version of the system that always displayed the same high degree of politeness, irrespective of whether the user showed no politeness.

2. Rationale

The above-reported literature strongly shows that communication with SDS is part of everyday life. Politeness theory has introduced the concept of face threats, e.g., affronts to a person's autonomy and/or needs of belonging. These face threats are mitigated via communicational behavior, such as hedges [45] and relativizing words.

Empirical studies indicate that SDS are evaluated on social dimensions comparably to humans. Those experimental studies have also shown that word usage impacts this evaluation.

The following study manipulates SDS word usage with regard to politeness. We conducted a 1x2 experimental design, where politeness was contrasted with rudeness. We define rudeness as a deliberate attack on the addressee's face that can be used both playfully and aggressively [46]. In this respect, rudeness is distinct from a mere lack of politeness when uttering a necessary face threat (i.e., a bald/on record strategy in the terms of politeness theory), as well as an unintentional face threat (e.g., an SDS asking a strongly religious user a question about a topic that is offensive to them) but consists of intentionally aggravating behavior. The interpretation often is dependent on the context (e.g., [47]).

We formulate three hypotheses that address participants' evaluation of both the social aspects of the SDS and its competence. Furthermore, by directly asking participants to suggest changes in SDS formulations, a direct measure on the word level was operationalized.

H1: A polite SDS will be judged as more likable and polite than will a rude SDS and its responses as more appropriate and pleasant than those from a rude SDS. A polite SDS will be more strongly perceived as a social interaction partner than a rude SDS.

H2: A polite SDS will be judged as more competent and trustworthy than a rude SDS.

H3: Rude responses will be perceived as more serious face threats than polite responses and will lead to more revisions than polite ones.

In the following, methods and materials are described.

3. Methods and Materials

In order to comply with the requirements of open science and to achieve transparency, we report how we determined our sample size, all (if any) data exclusions, all manipulations, and all measures in the study [48].

3.1. Participants. We recruited participants at an open house event that our university holds yearly for potential new students. We planned to recruit as many participants as possible, aiming for about 80. This would yield a power of about 80% in finding medium to large effects, as previous studies on similar phenomena have found.

In total, 58 persons participated in the study (35 females). The age of the participants ranged from 15 to 20 years old and the mean age was 16.91 ($SD = 1.08$) years old. All were native German speakers or spoke German since their early childhood. In Germany, students that aim for university entrance qualifications usually choose two intensive courses during the last two years of school. In our sample the most common choices were biology (36% of participants), English (33%), German (29%), and mathematics (21%). They reported to use a computer on average of 9.84 hours per week ($SD = 9.42$) and to use the Internet on average of 25.34 hours per week ($SD = 21.17$). Also, 86% of participants considered themselves to possess intermediate or advanced computer knowledge. Of all participants, 12% reported that they use SDS on a daily basis or several times a week, 8.6% reported to use SDS several times a month, and 77.6% reported that they rarely or never use SDS. Overall, 66% of participants indicated that they use SDS that are implemented in smartphones, like Siri and Google Now.

Experimental conditions did not differ for each of the above-reported descriptive variables, all $F_s > 0.832$, $p_s > .366$. Hence, these variables were not considered further on. No data were removed from the analysis. Participants received no compensation.

3.2. Materials. The study was conducted at a German university; all materials were in German. Materials and stimuli are available at <https://sites.google.com/site/sdspoliteness>. The examples we present in this paper are translations of the original materials. The speech was created using a text-to-speech synthesis tool available on Apple Mac computers. It was a female voice.

We told participants that we built/designed and trained an SDS to provide answers to university freshmen's questions. We stated that we therefore wanted them to assess the SDS's answers. The participants were told that the SDS was part of a mobile app provided to new psychology students to help them find their way around the university and their courses. To this end, users could ostensibly ask the SDS questions about

Frequently Asked Questions– FAQ – *From students, for freshmen.*

Planning of Studies:

- Do I have to take the exam in Social Psychology in the second semester?
- What are the consequences if I do not pass an exam in the scheduled window of time?
- How many internships do I have to complete?
- ....

Courses:

- How are Mrs Ralamo's teaching skills?
- Are Statistics really so difficult?
- Is there a compulsory attendance for courses?
- ....

FIGURE 1: Screenshot of the question list shown to participants.

different topics related to university life and the psychology program and receive informative answers. In reality, the app does not exist and participants were debriefed at the end of the experiment.

Participants listened to six answers presented by our own designed SDS, called ACURI. The answers addressed six typical questions of university freshmen. The questions were presented on a screen and after clicking on the respective question, the answer was presented as a sound file. The questions were the same in both experimental conditions, but the answers given by the SDS differed. In the *polite condition* the messages were phrased politely (e.g., “You can choose whether you want to...”). In the *rude condition* the face threats were strong(er), resulting in a rudely phrased message (e.g., “No, you have to do...”). Questions and answers as well as the original versions in German are shown in Table 1; see Table 2 for an example.

3.3. Procedure. Participants were tested in groups of five or six. Each member of the group was assigned the same experimental condition. A 1x2 between-subject design was realized, with $n = 29$ participants in each of the two conditions (polite versus rude). All participants were seated individually in front of a laptop computer and were provided with headphones. They received sheets containing information on the experimental procedure as well as a declaration of consent and data privacy. They also received a booklet with the questionnaire to be completed as part of the study. Our local ethics committee approved the study.

After reading the information and signing the form, the students' questions were presented on the computer screen. The participants were instructed to click on the questions to hear the SDS's responses. The responses were played as audio over the headphones. Participants were not free to choose which question to click; there was a fixed order in which the answers to the questions (sound files) were available (see Figure 1 for an example of a participants' screen).

After every response from the SDS, the participants were asked to turn a page of the booklet and respond to a number

of questions concerning the response they just heard. After all responses had been heard, the participants were asked to rate the SDS per se and provide demographic information. After completing the questionnaire, the participants were debriefed. The session took about 30 minutes.

3.4. Dependent Measures. Participants were asked to rate every single response of ACURI on the perceived face threat as well as the holistic impression of ACURI at the end of the survey.

3.4.1. Evaluation of the Responses. The participants rated each of the six responses on the following.

Pleasantness and Appropriateness. Participants were asked to indicate on 7-point bipolar semantic differentials whether they found the response (1) pleasant–unpleasant and (2) appropriate–inappropriate.

Perceived Face Threat Scale (PFT; [49]). The scale was originally designed to rate utterances in workplace conversations and later complaints in interpersonal contexts, regarding how much they threatened positive and negative face aspects. The authors found that responses on the scale differed depending on the type of the complaint, such that complaints that had focused on the disposition of the receiver were judged as more face-threatening. This makes the scale adequate for our goals. We are unaware of any other measure for evaluating face threat.

Sample items are “My partner's actions showed disrespect toward me” for threatening positive face and “My partner's actions constrained my choices” for threatening negative face. Because we did not pose a hypothesis regarding the different face aspects, we averaged scores on all 14 items into a single value. For the current study, all items were translated into German and adapted to collect ratings on “ACURI” instead of “my partner”. Participants responded on a 7-step Likert scale ranging from “not at all” to “very much”. Scale reliability was good with Cronbach's $\alpha = .91$.

TABLE 1: Student's questions and polite and rude phrasings of face threats in ACURI's responses.

Student Question	Polite Phrasing in ACURI's Response	Rude Phrasing in ACURI's RESPONSE
Do I have to take the exam in Social Psychology in the second semester?	You can choose whether you want to do the exam in your second or third semester. You find this information on the website if you look for the link to syllabus and exam information.	No, you have to do the social psychology exam in the second or third semester. You find this information on the website if you look for the link to syllabus and exam information.
What are the consequences if I do not pass an exam in the scheduled window of time?	That can happen but it's regrettable. Other courses can only be taken once the exam is passed. For your Bachelor's degree this is around 12 weeks. The exact number of hours varies from Bachelor to Master degree. If you get a chance take a look at the syllabus. After completing your internships you will for sure have a good idea about the different fields Psychologists work in.	Well, hard luck! You have to pass the exam before you can take other courses. As many as you need in order to get an idea of the different fields Psychologists work in. The exact number of hours varies from Bachelor to Master degree. To find out you will have to look into the syllabus.
How many internships do I have to complete?		
How are Mrs. Ralamo's teaching skills?	Students she perceives to be motivated to work will pass her class rather easily. So you will get a long with her well if you do the work in class.	Well it's going to be difficult for lazy ones. Make an effort and do the work in class and you won't have problems with her.
Are Statistics really that difficult?	If you make an effort you will do okay. In addition to that there is a statistics tutorial that you can enroll in.	People asking such a question will probably have problems. If you are already worried that you won't manage it'll probably turn out this way. Enroll in the statistics tutorial right away; the tutors regularly make miracles happen.
Is there a compulsory attendance for courses	For most lectures and classes attendance is not mandatory but you should try to go nonetheless because contents to be learned usually stick better this way then through self-study.	No, attendance is not mandatory for most lectures and classes. But how do you expect to learn if you don't go to class?

TABLE 2: Example of a student's question (SQ) and the polite and rude answers, respectively, of the spoken dialogue system named ACURI.

	Polite	Rude
SQ1	Do I need to take the social psychology exam in the second semester?	
ACURI	You can choose whether you want to do the exam in your second or third semester. . . .	No, you have to do the social psychology exam in the second or third semester. . . .

3.4.2. *Evaluation of the SDS.* The participants evaluated the SDS regarding several subjective appraisals, epistemic trust, and whether they perceived it as a social agent.

Subjective Assessment of Speech System Interfaces. The participants rated the SDS on the Subjective Assessment of Speech System Interfaces measure (SASSI, [50]). This instrument was developed as an extension of earlier measures to evaluate the usability of graphic interfaces and focuses on subjective aspects of SDS usability. It consists of six subscales: response accuracy (nine items, e.g., “the system makes few errors”), likability (nine items, e.g., “the system is pleasant”), cognitive demand (five items, e.g., “a high level of concentration is required when using the system”), annoyance (five items, e.g., “the interaction with the system is frustrating”), habitability (four items, e.g., “I was not always sure what the system was doing”), and speed (two items, e.g., “the interaction with the system is fast”). There are alternative measures focusing on SDS usability, most notably the CCIR-BT [51] with similar subscales. However, the SASSI measure has been more widely used in subsequent research (e.g., [52]).

In the current study, the participants indicated their agreement to the statements on 7-point Likert scales. Scale reliabilities for the response accuracy and likability subscales were good, with $\alpha = .84$ and $.87$, respectively. Reliabilities for the cognitive demand and annoyance subscales were acceptable with $\alpha = .68$ and $.72$, respectively. However, the subscale reliabilities for the habitability and speed subscales were inadequate with $\alpha = -.06$ and $.12$, respectively. This might be explained by the fact that the participants were not using the SDS themselves but merely judging the SDS's utterances. The two subscales were dropped from further analyses.

Hence, four measures, each one for a subscale of SASSI serves as subjective assessment of ACURI.

Perceiving the SDS as a Social Agent. The participants indicated how much they perceived the SDS as a social agent using a measure developed by Holtgraves, Ross, Weywadt, and Han [21]. This inventory was developed in order to assess whether users ascribe human-like qualities to chatbots. The measure consists of two subscales that measure perceptions of conversational skill (three items, e.g., “how engaging is the system?”) and pleasantness (three items, e.g., “how thoughtful is the system?”). The participants responded on 7-point semantic differentials (e.g., “not at all thoughtful–very thoughtful”). In the original publication, chatbots using the user's first name were found to be evaluated more positively on these scales. In our study, subscale reliabilities were good to satisfactory with Cronbach's $\alpha = .67$ and $.81$, respectively.

Epistemic Trust. The participants rated how much they trusted the SDS as a source of knowledge using the Münster Epistemic Trustworthiness Inventory (METI; [53]). This measure is based upon the ABI model mentioned in the introduction [24] and consists of 5-point bipolar adjective pairs. The subscales measure goodwill (four items, e.g., “moral–immoral”), expertise (six items, e.g., “qualified–unqualified”), and integrity (five items, e.g., “honest–dishonest”). All subscales exhibited satisfactory consistencies with Cronbach's $\alpha = .73$, $.81$, and $.60$, respectively.

There are alternative measures that measure similar constructs, such as the Credibility scales by McCroskey and Teven [54]. However, the METI instrument differs from these in that it explicitly focuses on *epistemic* trust, that is, whether the target of the evaluations is a trustworthy source for the knowledge that the user seeks. This was desirable for our research question.

Suggestions for Rephrasing. Participants were given the opportunity to rephrase ACURIs answers and to mention things that, from their perspective, should be changed. They provided their answers in writing as response to this instruction: “Please listen to the answer of the question once again. You may now note potential suggestions for changing the answer.”

For the analysis, we used a bottom-up, data-driven process to identify five categories of statements of what should be changed according to the participants: (1) apply strategies to mitigate FTAs (e.g., “use ‘it is recommended that you..’ instead of ‘you have to..’”); (2) formulate the utterance in a more direct and neutral way, e.g., “should just answer the question and not make a proposal”; (3) change the prosody or pronunciation, e.g., “a brighter and less monotonous voice would be better”; (4) provide more precise information, e.g., “give information about where the tutorial takes place”; and (5) no changes are necessary, e.g., “the hint was helpful”.

4. Results

We used the lme4 package [55] for the R statistical software and entered the six individual responses as a random effect and the politeness condition (polite/rude) as a fixed effect into linear mixed-effects models. For the ratings collected after the whole discourse, we calculated ANOVAs and MANOVAs, depending on the measure as reported below.

With *hypothesis 1*, we assumed that a polite SDS would be judged as more likable and polite responses as more appropriate and pleasant than a rude SDS. Furthermore, we expected a polite SDS to be more strongly perceived as a social interaction partner than a rude SDS. These expectations

were partly confirmed: both ratings of pleasantness and appropriateness measures yielded the expected effects. Polite responses were perceived as more appropriate ($M = 6.53$, $SE = 0.28$) than rude responses ($M = 5.59$, $SE = 0.28$), $F(1,55) = 13.59$, $p < .001$. Polite responses were also perceived as more pleasant ($M = 6.07$, $SE = 0.28$) than rude responses ($M = 4.97$, $SE = 0.29$), $F(1,55) = 10.43$, $p < .001$.

Contrary to our expectations, on the holistic level, the SASSI likability subscale did not show a significant effect for the politeness condition, $F(1,51) = 1.87$, $p = .177$. Both groups judged likability of ACURI as moderate (polite: $M = 4.62$, $SD = 1.07$; rude: $M = 4.18$, $SD = 1.19$).

Regarding how much participants perceived the SDS as a social agent with human-like properties, we used the measure by Holtgraves and colleagues [21]. The pleasantness subscale showed a significant effect of the politeness condition in the expected direction. As such, the polite SDS was also judged as more pleasant ($M = 5.67$, $SE = 0.23$) than the rude SDS ($M = 4.14$, $SE = 0.23$), $F(1,51) = 17.88$, $p < .001$. However, the polite SDS's conversational skill was not judged to be higher, $F(1,51) = 1.04$, $p = .31$.

With **hypothesis 2**, we assumed that a polite SDS would be judged as more competent and trustworthy than a rude SDS. This hypothesis was mostly confirmed. The polite SDS received higher ratings on the SASSI response accuracy subscale ($M = 4.65$, $SE = 0.13$) than the rude SDS ($M = 4.23$, $SE = 0.13$), $F(1,51) = 4.48$, $p = .039$. The MANOVA with the METI subscales showed a significant effect for the politeness condition, $F(3,49) = 4.15$, $p = .011$. The follow-up analyses showed that the polite SDS was judged as showing more goodwill (polite: $M = 3.63$, $SD = 0.72$; rude: $M = 2.90$, $SD = 0.78$) and integrity (polite: $M = 3.90$, $SD = 0.63$; rude: $M = 3.51$, $SD = 0.58$) but not more expertise (polite: $M = 3.70$, $SD = 0.71$; rude: $M = 3.54$, $SD = 0.73$) than the rude SDS.

With the **hypothesis 3**, we expected that rude responses would be perceived as more serious face threats than polite responses. H3 was confirmed. According to the PFT, rude responses were perceived as more face-threatening ($M = 3.22$, $SE = 0.14$) than polite responses ($M = 2.46$, $SE = 0.14$), $F(1,55) = 22.46$, $p < .001$.

The average amount of revision proposals regarding all respective answers was comparable in both conditions. On average, participants made suggestions on 2.48 ($SD = 2.13$) answers in the polite condition and on 2.62 answers in the rude condition ($SD = 1.97$; $F(1,56) = 0.07$, $p = .799$). However, the amount of suggestions regarding the categories differed depending on the respective response (see Table 3 for details). We calculated Fisher's exact tests for each response to test for differences between conditions. This is a statistical technique for the analysis of deviations from expectation in a contingency table [56] and is especially suited for smaller samples.

The condition had an influence on answers to SQ 2 ($\chi^2(4, N = 58) = 16.550$, $p < .001$), SQ4 ($\chi^2(4, N = 58) = 9.184$, $p = .029$), and SQ5 ($\chi^2(4, N = 58) = 12.025$, $p = .006$), but not on answers to SQ1 (SQ1: $\chi^2(4, N = 58) = 2.545$, $p = .725$, *ns.*), SQ3 ($\chi^2(5, N = 58) = 8.086$, $p = .110$, *ns.*), and SQ6 ($\chi^2(4, N = 58) = 5.463$, $p = .213$, *ns.*). In the rude condition, most changes were aimed at strategies to mitigate FTAs (29 % on

average). In the politeness condition, most suggestions were offered in categories 3 or 4. Participants referred to providing more precise information and on changing or enhancing the employed voice (both categories each around $M = 14\%$). These two categories were each chosen for 6% of the answers by participants of the rude condition.

5. Discussion

To sum up, the results of this study show that polite responses were perceived as more appropriate and pleasant than rude responses. Overall, the SDS was also perceived as more pleasant than the rude SDS. Both groups judged the likability of the SDS as moderate, with no differences between conditions. Also, conversational skill was not judged differently between conditions. The polite SDS was perceived as more accurate, showing more goodwill and integrity but not as having more expertise than the rude SDS. Rude responses were perceived as more face-threatening than polite responses. In the rude condition, most changes were aimed at strategies to mitigate FTAs (29 % on average), while, in the politeness condition, participants referred to providing more precise information and on changing or enhancing the employed voice.

Using a very clear and straightforward manipulation, we aimed to identify how the human principle of politeness is taken in consideration in SDS communication. In our setting, SDS answered to university freshmen's typical questions on student life. The only differences between our two experimental conditions were in the "how" part of the contribution, with rather rude or polite answers. The results mostly mirror those obtained with judgments of humans [12, 30]. These studies suggest that although both content and social information are transmitted with the same words, recipients seem to distinguish these two aspects; we routinely tease apart the "how" and "what" part of a contribution. In both previous studies and our present study, almost all social judgments, such as likability or the goodwill aspect of trustworthiness, were evaluated more positively in the polite condition. However, neither found evidence for differences in judgments of expertise, as a more content-related aspect of trustworthiness. However, in our study the accuracy of the system, which is also an arguably content-related aspect, was judged as higher for the polite system. Future research should ascertain whether this is a specific aspect of communication with an SDS or maybe a consequence of the specific realization of our conditions.

The results also indicate that advances on the technological side are expectations on the users' side. Politeness and adaptive communication require competent systems. Our between-subject design, which did not provide a direct comparison of rude and polite behavior, shows clearly that an SDS is judged relatively strictly according to its communication behavior. The users do not seem to be willing to be lenient simply because the system is not a human.

Our study has some limitations: one is that we had relatively young users assess the SDS. While those do represent typical users and the whole setting was ecologically valid for them, we cannot transfer the SDS evaluation results to other groups with less experience in technology (see the

TABLE 3: Change suggestions for each of the five categories from participants for each of the six SDS answers according to conditions (polite versus rude), reported in percentage.

category	Answer 1		Answer 2		Answer 3		Answer 4		Answer 5		Answer 6		Total	
	polite	rude	polite	rude	polite	rude	polite	rude	polite	rude	polite	rude	polite	rude
1 (mitigate FTA)	10	21	14	59	7	10	3	28	7	41	10	17	9	29
2 (direct)	3	0	3	0	3	0	0	0	3	0	10	10	4	2
3 (voice)	28	31	7	3	17	0	10	3	7	0	14	0	14	6
4 (precise)	7	3	21	0	21	21	14	3	21	7	3	0	15	6
5 (no change)	52	45	55	38	52	69	72	66	62	52	62	72	59	57

digital natives debate; [57]). Second, we simulated an indirect communication setting. Our participants did not interact with the SDS themselves; instead, they listened to responses of the SDS. There is evidence that a direct engagement with the SDS produces different results and absorbs some of the analytic abilities participants have shown in their evaluation of our ACURI [58].

One practical implication of this study might be drawn from our empirical evidence: regarding the first impression of an SDS and the evaluation of its acceptance and trustworthiness, the wording used by the SDS seems to have a considerable impact. Hence it might be worth engaging in creating more flexible and human-like technology. Communication principles such as politeness and alignment provide straightforward assumptions that might be embedded in technology and tested in experimental settings. Although the more attempts that are aimed at making assistants like ACURI become more human-like in their communication style, it is possible that such systems might also enhance miscommunication and misunderstanding. As an example, an SDS that is programmed to offer polite and indirect communication might produce responses that leave more room for interpretation and are thus less clear. In order to competently implement social communication factors in SDS, designers need to be aware how these principles come into play in different communication contexts. In this way, research into the social factors of human communication is highly relevant for the field of human-computer interaction.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by a grant awarded to the second author within the German Research Foundation's (DFG) Research Training Group GRK 1712: *Trust and Communication in a Digitized World*.

References

- [1] N. Mavridis, "A review of verbal and non-verbal human-robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, no. 1, pp. 22–35, 2015.
- [2] R. López-Cózar, Z. Callejas, D. Griol, and J. F. Quesada, "Review of spoken dialogue systems," *Loquens*, vol. 1, no. 2, Article ID e012, 2014.
- [3] F. Manjoo, "The Echo from Amazon brims with groundbreaking promise," <http://www.nytimes.com/2016/03/10/technology/the-echo-from-amazon-brims-with-groundbreaking-promise.html>, 2016.
- [4] M. F. McTear, *Spoken Dialogue Technology: toward The Conversational User Interface*, Springer Science & Business Media, 2004.
- [5] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- [6] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, *In the Shades of the Uncanny Valley: An Experimental Study of Human-Chatbot Interaction*, Future Generation Computer Systems, 2018.
- [7] N. Dethlefs, H. Hastie, H. Cuayáhuil, Y. Yu, V. Rieser, and O. Lemon, "Information density and overlap in spoken dialogue," *Computer Speech Language*, vol. 37, pp. 82–97, 2016.
- [8] O. Vinyals and Q. Le, "A neural conversational model," arXiv preprint, arXiv:1506.05869. ISO 690, 2015.
- [9] G. A. Linnemann and R. Jucks, "As in the question, so in the answer? Language style of human and machine speakers affects interlocutors' convergence on wordings," *Journal of Language and Social Psychology*, vol. 35, no. 6, pp. 686–697, 2016.
- [10] C. I. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, MIT Press, Cambridge, UK, 2005.
- [11] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson, "Towards human-like dialogue systems," *Speech Communication*, vol. 50, no. 8, pp. 630–645, 2008.
- [12] R. Jucks, G. A. Linnemann, F. M. Thon, and M. Zimmermann, "Trust the words: insights into the role of language in trust building in a digitalized world," in *Trust and Communication in a Digitized World*, B. Blöbaum, Ed., pp. 225–237, Springer International Publishing, 2016.
- [13] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue," *AI Magazine*, vol. 22, no. 4, pp. 39–51, 2001.

- [14] R. E. Mayer, W. L. Johnson, E. Shaw, and S. Sandhu, "Constructing computer-based tutors that are socially sensitive: politeness in educational software," *International Journal of Human-Computer Studies*, vol. 64, no. 1, pp. 36–42, 2006.
- [15] K. Porayska-Pomsta, M. Mavrikis, and H. Pain, "Diagnosing and acting on student affect: the tutors perspective," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1, pp. 125–173, 2008.
- [16] M. Tai, I. Arroyo, and B. P. Woolf, "Teammate relationships improve help-seeking behavior in an intelligent tutoring system," in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 239–248, Springer, Berlin, Heidelberg, 2013.
- [17] B. Whitworth, "Politeness as a social software requirement," *International Journal of Virtual Communities and Social Networking*, vol. 1, no. 2, pp. 65–84, 2009.
- [18] B. M. McLaren, K. E. DeLeeuw, and R. E. Mayer, "Polite web-based intelligent tutors: can they improve learning in classrooms?" *Computers & Education*, vol. 56, no. 3, pp. 574–584, 2011.
- [19] P. Brown and S. C. Levinson, *Politeness, Some universals in language Usage*, Cambridge University Press, Cambridge, UK, 1987.
- [20] B. Brummernhenrich and R. Jucks, "Managing face threats and instructions in online tutoring," *Journal of Educational Psychology*, vol. 105, no. 2, pp. 341–350, 2013.
- [21] T. Holtgraves, S. Ross, C. Weywadt, and T. Han, "Perceiving artificial social agents," *Computers in Human Behavior*, vol. 23, no. 5, pp. 2163–2174, 2007.
- [22] A. De Angeli, W. Gerbino, E. Nodari, and D. Petrelli, "From tools to friends: where is the borderline?" in *Proceedings of the UM99 Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, pp. 1–10, Springer, Berlin, Germany, 1999.
- [23] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction," *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, pp. 171–181, 2001.
- [24] R. C. Mayer, J. H. Davis, and F. D. Shoorman, "An intergration model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.
- [25] S. Tseng and B. J. Fogg, "Credibility and computing technology," *Communications of the ACM*, vol. 42, no. 5, pp. 39–44, 1999.
- [26] L. Hoffmann, N. C. Krämer, A. Lam-chi, and S. Kopp, "Media equation revisited: do users show polite reactions towards an embodied agent?" in *Intelligent Virtual Agents*, Z. Ruttkay, M. Kipp, A. Nijholt, and H. H. Vilhjálmsson, Eds., pp. 159–165, Springer, Berlin, Germany, 2009.
- [27] A. De Angeli and S. Brahmam, "I hate you! Disinhibition with virtual partners," *Interacting with Computers*, vol. 20, no. 3, pp. 302–310, 2008.
- [28] J. Culpeper, "Towards an anatomy of impoliteness," *Journal of Pragmatics*, vol. 25, no. 3, pp. 349–367, 1996.
- [29] C. Nass, "Etiquette equality: exhibitions and expectations of computer politeness," *Communications of the ACM*, vol. 47, no. 4, pp. 35–37, 2004.
- [30] B. Brummernhenrich and R. Jucks, "'He shouldn't have put it that way!' How face threats and mitigation strategies affect person perception in online tutoring," *Communication Education*, 2015.
- [31] R. Jucks, L. Päuler, and B. Brummernhenrich, "'I need to be explicit: you're wrong': impact of face threats on social evaluations in online instructional communication," *Interacting with Computers*, vol. 28, no. 1, pp. 73–84, 2016.
- [32] S. L. Jessmer and D. Anderson, "The effect of politeness and grammar on user perceptions of electronic mail," *North American Journal of Psychology*, vol. 3, no. 2, pp. 331–346, 2001.
- [33] L. Dybkjær and N. O. Bernsen, "Usability issues in spoken dialogue systems," *Natural Language Engineering*, vol. 6, pp. 243–271, 2000.
- [34] D. H. McKnight, "Trust in information technology," in *The Blackwell Encyclopedia of Management*, B. G. Davis, Ed., vol. 7 of *Management Information Systems*, pp. 329–331, Blackwell, Malden, MA, USA, 2005.
- [35] R. C. Mayer and J. H. Davis, "The effect of the performance appraisal system on trust for management: a field quasi-experiment," *Journal of Applied Psychology*, vol. 84, no. 1, pp. 123–10, 1999.
- [36] D. H. McKnight and N. L. Chervany, "Trust and distrust definitions: one bite at a time," in *Trust in Cyber-Societies*, R. Falcone, M. Singh, and Y.-H. Tan, Eds., pp. 27–54, Springer, Berlin, Heidelberg, 2001.
- [37] M. D. Pickard, J. K. Burgoon, and D. C. Derrick, "Toward an objective linguistic-based measure of perceived embodied conversational agent power and likeability," *International Journal of Human-Computer Interaction*, vol. 30, no. 6, pp. 495–516, 2014.
- [38] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers," *Journal of Pragmatics*, vol. 42, no. 9, pp. 2355–2368, 2010.
- [39] C. Torrey, A. Powers, M. Marge, S. R. Fussell, and S. Kiesler, "Effects of adaptive robot dialogue on information exchange and social relations," in *Proceedings of the HRI 2006: 2006 ACM Conference on Human-Robot Interaction*, pp. 126–133, USA, March 2006.
- [40] J. J. Bradac, A. Mulac, and A. House, "Lexical diversity and magnitude of convergent versus divergent style shifting: perceptual and evaluative consequences," *Language Communication*, vol. 8, no. 3, pp. 213–228, 1988.
- [41] S. Gupta, M. Walker, and D. Romano, "How rude are you?: evaluating politeness and affect in interaction," *Affective Computing and Intelligent Interaction*, pp. 203–217, 2007.
- [42] M. A. Walker, J. E. Cahn, and S. J. Whittaker, "Improvising linguistic style: social and affective bases for agent personality," in *Proceedings of the First International Conference on Autonomous Agents*, pp. 96–105, 1997.
- [43] A. C. Graesser, "Learning, thinking, and emoting with discourse technologies," *American Psychologist*, vol. 66, no. 8, pp. 746–757, 2011.
- [44] M. De Jong, M. Theune, and D. Hofs, "Politeness and alignment in dialogues with a virtual guide," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2008*, pp. 206–213, Portugal, May 2008.
- [45] M. Thiebach, E. Mayweg-Paus, and R. Jucks, "'Probably true' says the expert: how two types of lexical hedges influence students' evaluation of scientificness," *European Journal of Psychology of Education*, vol. 30, no. 3, pp. 369–384, 2015.
- [46] D. Bousfield and J. Culpeper, "Impoliteness, eclecticism and diaspora," *Journal of Politeness Research*, vol. 4, no. 2, pp. 161–168, 2008.
- [47] N. Vergis and M. Terkourafi, "The role of the speakers emotional state in im/politeness assessments," *Journal of Language and Social Psychology*, vol. 34, no. 3, pp. 316–342, 2015.

- [48] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011.
- [49] W. R. Cupach and C. L. Carson, "Characteristics and consequences of interpersonal complaints associated with perceived face threat," *Journal of Social and Personal Relationships*, vol. 19, no. 4, pp. 443–462, 2002.
- [50] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," *Natural Language Engineering*, vol. 6, no. 3, pp. 287–303, 2000.
- [51] L. B. Larsen, "Assessment of spoken dialogue system usability—what are we really measuring?" in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.
- [52] S. Möller, P. Smeele, H. Boland, and J. Krebber, "Evaluating spoken dialogue systems according to de-facto standards: a case study," *Computer Speech & Language*, vol. 21, no. 1, pp. 26–53, 2007.
- [53] F. Hendriks, D. Kienhues, and R. Bromme, "Measuring laypeople's trust in experts in a digital age the Muenster Epistemic Trustworthiness Inventory (METI)," *PLoS ONE*, vol. 10, no. 10, Article ID e0139309, 2015.
- [54] J. C. McCroskey and J. J. Teven, "Goodwill, a reexamination of the construct and its measurement," *Communication Monographs*, vol. 66, no. 1, p. 90, 1999.
- [55] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, 2015.
- [56] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, pp. 131–153, 1992.
- [57] F. Salajan, D. Schonwetter, and B. Cleghorn, "Student and faculty inter-generational digital divide: fact or fiction?" *Computers and Education*, vol. 53, no. 3, pp. 1393–1403, 2010.
- [58] G. A. Linnemann and R. Jucks, "Can I trust the spoken dialogue system because it uses the same words as i do?—influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction," *Interacting with Computers*, pp. 173–186, 2018.