Influence of Time and Risk on Response Acceptability in a Simple Spoken Dialogue System

Andisheh Partovi and Ingrid Zukerman
Clayton School of Information Technology, Monash University
Clayton, Victoria 3800, Australia
Andi.Partovi@monash.edu, Ingrid.Zukerman@monash.edu

Abstract

We describe a longitudinal user study conducted in the context of a Spoken Dialogue System for a household robot, where we examined the influence of time displacement and situational risk on users' preferred responses. To this effect, we employed a corpus of spoken requests that asked a robot to fetch or move objects in a room. In the first stage of our study, participants selected among four response types to these requests under two risk conditions: low and high. After some time, the same participants rated several responses to the previous requests — these responses were instantiated from the four response types. Our results show that participants did not rate highly their own response types; moreover, they rated their own response types similarly to different ones. This suggests that, at least in this context, people's preferences at a particular point in time may not reflect their general attitudes, and that various reasonable response types may be equally acceptable. Our study also reveals that situational risk influences the acceptability of some response types.

1 Introduction

Spoken Dialogue Systems (SDSs) must often engage in follow-up interactions to deal with Automatic Speech Recognizer (ASR) errors or elucidate ambiguous or inaccurate requests (which are exacerbated by ASR errors):

- ASR errors, although significantly reduced in recent times,[1] may produce wrong entities or actions, or ungrammatical utterances that cannot be processed by a Spoken Language Understanding (SLU) system (e.g., "the plate inside the microwave" being misheard as "*of plating sight* the microwave").[2]
- People often express themselves ambiguously or inaccurately (Trafton et al., 2005; Moratz and Tenbrink, 2006; Funakoshi et al., 2012; Zukerman et al., 2015). An ambiguous reference to an object matches several objects well, while an inaccurate reference matches one or more objects partially. For instance, a reference to a "big blue mug" is ambiguous if there is more than one big blue mug, and inaccurate if there are two mugs – one big and red, and one small and blue.

In the last two decades, research in response generation has focused on techniques that generate response policies that optimize dialogue completion, using *Markov Decision Processes* (*MDPs*), e.g., (Singh et al., 2002; Lemon, 2011), and *Partially Observable MDPs* (*POMDPs*), e.g., (Williams and Young, 2007; Gašić and Young, 2014). Recently, deep-learning algorithms have been used to generate dialogue responses on the basis of request-response pairs, e.g., (Li et al., 2016; Prakash et al., 2016; Serban et al., 2017). Human and simulation-based evaluations of MDP and POMDP systems focus on dialogue completion, while evaluations of deep-learning algorithms focus on individual responses.

In this paper, we draw inspiration from research in Recommender Systems, where Amatriain et al. (2009) and Said and Bellogín (2018) showed that over time, users gave inconsistent ratings to items, leading to the "magic barrier" to prediction accuracy in Recommender Systems (Said and Bellogín, 2018). This prompted us to posit that people may also be inconsistent when assessing responses in a dialogue at different times, which may affect the results of human evaluations.

To investigate this claim, we conducted a longitudinal study in the context of an SDS for a household robot. We first collected a corpus of spoken requests that asked a robot to fetch or move

[1] 9to5google.com/2017/06/01/google-speech-recognition-humans/.

[2] All the sample ASR outputs in this paper are real.

Proceedings of the SIGDial 2019 Conference, pages 307–319
Stockholm, Sweden, 11-13 September 2019.

objects in a room. Our participants were shown the top ASR outputs for these requests (the intention was to replicate the information available to an SDS, without the extra information people can glean from what they hear). They were also told that these requests had to be executed under two risk conditions: *low risk*, where the consequences of performing the wrong action are trivial, and *high risk*, where performing the wrong action could significantly inconvenience the speaker. The participants had to choose among four response types: DO the request without further interaction, CONFIRM the intended object, ask the requester to CHOOSE between a few candidate objects, or ask the requester to REPHRASE all or part of the request. After 1.5-2 years, the same participants were shown the original requests and ASR outputs, and were asked to rate responses generated from their previously selected response types and from other sources, in particular response types selected by one of the authors and by a classifier trained on the author's chosen response types.

Our findings show that (1) participants downrated responses sourced from their previously chosen response types; and (2) these responses were liked as much as *different* responses sourced from the response types selected by one of the authors or by the above-mentioned classifier. The first result indicates that, at least in the context of one-shot dialogues with an SDS for a household robot, people's preferred response types at a particular point in time may not reflect their general attitudes. The second result suggests that, instead of one best response type, several reasonable response types may be acceptable, including those selected by a classifier trained on a non-target but relevant corpus.

We also investigated the influence of situational risk on the acceptability of response types. We found that (3) as expected, under the high-risk condition, the preferred response types were generally more conservative than under the low-risk condition; but (4) surprisingly, participants' attitudes toward certain response types, e.g., CONFIRM, were not affected by risk.

The rest of this paper is organized as follows. In the next section, we discuss related work. Our experimental setup is described in Section 3. In Section 4, we present our classifier and the features used to train it. The results of our experiment are described in Section 5, and concluding remarks appear in Section 6.

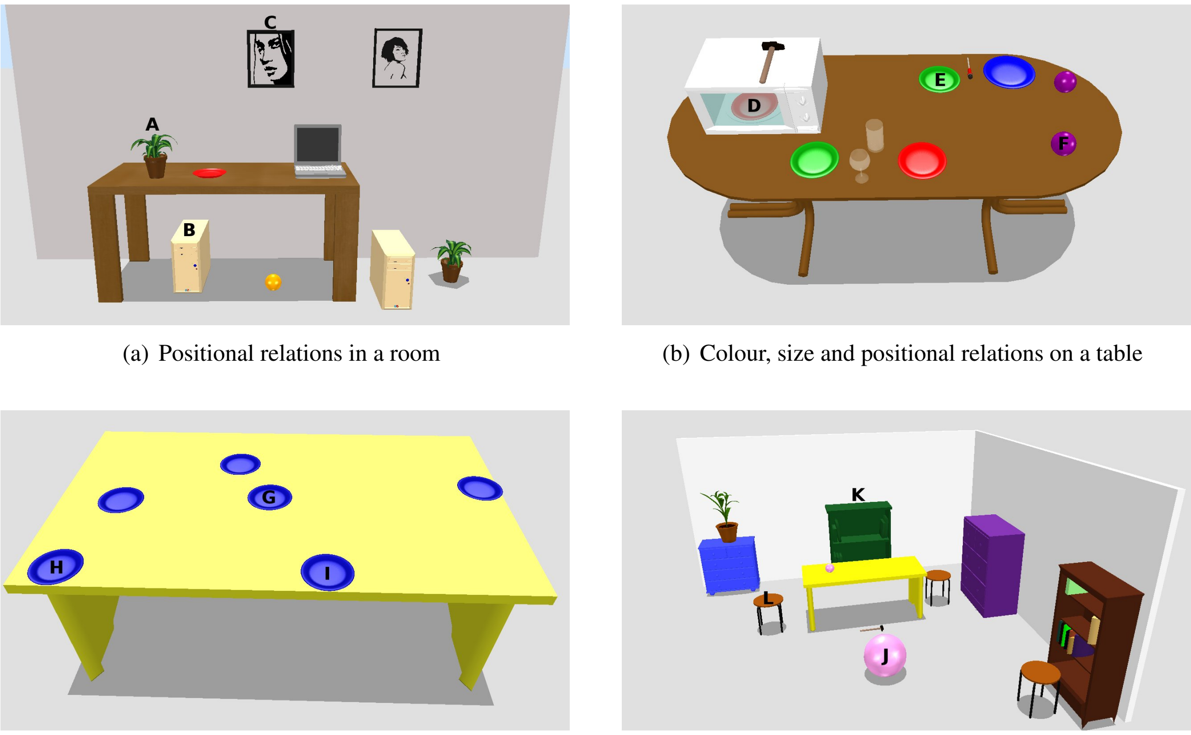
2 Related Work

Decision-theoretic approaches have been the accepted standard for response generation in dialogue systems for some time (Carlson, 1983). These approaches were initially implemented in SDSs as Bayesian reasoning processes that optimize a system's confidence when making myopic (one-shot) decisions regarding dialogue acts (Paek and Horvitz, 2000; Sugiura et al., 2009), and as *Dynamic Decision Networks* that make decisions about dialogue acts over time (Horvitz et al., 2003; Liao et al., 2006).

MDPs (Singh et al., 2002; Lemon, 2011), *POMDPs* (Williams and Young, 2007; Gašić and Young, 2014), and their extensions *Hidden Information State Model* (Young et al., 2010, 2013) and *Conversational Entity Dialogue Model* (Ultes et al., 2018) were used, often in combination with *Reinforcement Learning* (*RL*), to learn policies that optimize dialogue completion on the basis of feedback given by real or simulated users.

Recently, deep learning has been applied to various aspects of SDSs (Wen et al., 2015; Li et al., 2016; Mrkšic et al., 2017; Prakash et al., 2016; Serban et al., 2017; Tseng et al., 2018; Yang et al., 2017). Wen et al. (2015) and Tseng et al. (2018) considered the generation of linguistically varied responses; Li et al. (2016) and Prakash et al. (2016) produced dialogue contributions of chatbots; and Serban et al. (2017) generated helpdesk responses and Twitter follow-up statements. Mrkšic et al. (2017) proposed a dialogue-state tracking framework, and Yang et al. (2017) a mechanism for slot tagging and user-intent and system-action prediction in slot-filling applications. A combination of deep learning and RL has been used in end-to-end dialogue systems that query a knowledge-base, where user utterances are mapped to a clarification question or a knowledge-base query (Williams and Zweig, 2016; Zhao and Eskenazi, 2016; Dhingra et al., 2017). All these systems harness large corpora comprising request-response pairs to learn responses that are assumed to be better than alternative options.

Like evaluations based on simulated users, human evaluations of (PO)MDP/RL systems focus on successful dialogue completion (Singh et al., 2002; Thomson et al., 2008; Young et al., 2010), while human evaluations of deep-learning systems assess individual responses (Wen et al., 2015; Li et al., 2016; Prakash et al., 2016; Serban et al., 2017; Dhingra et al., 2017).



(a) Positional relations in a room

(b) Colour, size and positional relations on a table

(c) Projective and positional relations on a table

(d) Colour, size and positional relations in a room

Figure 1: Household scenes used in our study

The findings reported in this paper contribute to (PO)MDP/RL research by determining whether there are factors other than dialogue completion that affect the suitability of responses, and to deep-learning research by ascertaining whether indeed there is a single best response to each request.

The research described in (Jurčíček et al., 2011) and (Liu et al., 2016) shed light on ancillary aspects of human evaluations of system responses. The former compared evaluations by Amazon Mechanical Turk workers with evaluations by participants recruited for a lab experiment; and the latter conducted user studies to determine the validity of word-based evaluation metrics.

This paper also addresses ancillary aspects of human response evaluations, viz the influence of temporal displacement and situational risk on users' attitudes toward response types, and users' opinions of response types obtained from different sources (including a classifier trained on a corpus that differs from the target corpus).

3 Experimental Setup

Our experiment comprises two main stages: (1) responding to requests, and (2) rating responses to the same requests.

Creating a corpus of requests

We created a corpus of requests by collecting a corpus of spoken descriptions, and converting them to requests.

To collect the spoken descriptions, we replicated the experiment described in (Zukerman et al., 2015), but we used the Google ASR, instead of the Microsoft Speech API. In our experiment, the top-ranked outputs produced by this ASR had a 13% word error rate, which resulted in 53% of the descriptions having imperfect top-ranked ASR outputs. In addition, 33% of the descriptions had errors in all top four ASR outputs.

Following the protocol in (Zukerman et al., 2015), 35 participants were asked to describe 12 designated objects (labeled **A** to **L**) in four scenes (Figure 1); speakers were allowed to restate the description of an object up to two times. In total, we recorded 478 descriptions such as the following: "the flower on the table" (object **A** in Figure 1(a)), "the plate inside the microwave" (object **D** in Figure 1(b)), "the plate at the center of the table" (object **G** in Figure 1(c)), and "the large pink ball in the middle of the room" (object **J** in Figure 1(d)). 20% of the descriptions had an unintelligible object in all ASR outputs, e.g., "the *Heartist* under the table", 17.9% were ambiguous (several objects matched the description), and only 3.8% were inaccurate (no object matched the description perfectly).

We retained 292 descriptions,[3] and for each description, we used the top four ASR outputs. The corpus of requests, denoted *RequestCorpus*, was created by prefixing the verb "get" (for small objects) or "move" (for large objects) to each ASR output (which remained unchanged), e.g., "*get* the flower on the table". This corpus was divided into sets of at most 12 requests (one request per object, mostly from one speaker).

Demographic and risk-propensity information

We gathered information about the participants' gender, English nativeness, age, education and risk propensity. For the last item, we showed the participants twelve statements obtained from (Rohrmann, 2005): six risk-proneness statements, e.g., "I follow the motto 'nothing ventured, nothing gained' ", and six risk-aversion statements, e.g., "My decisions are always made carefully and accurately"; (dis)agreement was indicated on a 1-5 Likert scale. The hope was that these information items would assist in predicting participants' responses.

Stage 1 – Responding to requests

This corpus was collected through an online survey where participants had to indicate how they would respond to potentially misheard requests. Each participant was shown at most 12 requests from *RequestCorpus* (spoken by other people). Each request consisted of four verb-prefixed ASR outputs, and was accompanied by a version of the appropriate image in Figure 1 where the objects were numbered (to enable participants to identify any object as the referent). Each participant was then asked to select one of four response types for each request: DO, CONFIRM, CHOOSE or REPHRASE. Figure 3 in Appendix A displays a screenshot containing a numbered version of Figure 1(a), four ASR outputs for a request for object #5 (labeled **B** in Figure 1(a)), and the four response types.

Prior to presenting the survey questions, participants were given a training example containing the descriptions shown below in italics:

DO: *Fetch object # [This response is suitable if you are sure which object you should get].* Here participants were asked to enter the number of the object they would get or move.

CONFIRM: *Ask: Did you mean object #? [This response is suitable if you feel the need to confirm the requested object before taking action].* Here too participants were asked to enter the number of the object they were confirming.

CHOOSE: *Ask: Which object did you mean? [This response is suitable when you are hesitating between several objects].* In this case, participants were asked to enter the numbers corresponding to their candidate objects.

REPHRASE: *Ask: Please rephrase your request. [This response is suitable when a request is so garbled you can't understand it].*[4]

These choices were made under two risk conditions: *low risk* – where participants were told that the requested object must be delivered to someone in the same room; and *high risk* – where they were told that the object must be delivered to a remote location (Figure 3). These settings were designed to discriminate between situations where mistakes are fairly inconsequential and situations where mistakes are costly.

40 people took part in this stage of the experiment, but six dropped out after this stage. Half of the remaining participants were male, and 18 were native English speakers. 4 participants were between 18-24 years of age, 16 between 25-34 years of age, 7 between 35-44, and 7 over 45. In terms of education, 5 participants had a secondary education, 16 had a Bachelor, 8 a Masters, and 5 a PhD. To assess the participants' risk propensity, we subtracted their total risk-aversion score from their total risk-proneness score (the total risk-aversion/proneness score was calculated by adding up the Likert score of the six risk-aversion/proneness statements): 16 participants were risk prone, 8 were risk averse, and 10 were fairly neutral (the difference between the scores was less than 3).

In total, this corpus, denoted *ResponseCorpus*, contains 584 response types (= 292 requests × 2 conditions), which are distributed as shown in Columns 2 and 3 of Table 1.

To determine the influence of speaker diversity on classifier performance (Section 4), we created a second corpus, denoted *AuthorCorpus*, where one of the authors selected response types for all the

[3] 186 descriptions were removed as follows: 20 and 45 descriptions that were not tagged by Stage 1 and Stage 2 participants respectively, 59 descriptions that could not be processed by the SLU system, and 62 descriptions that had more than one prepositional phrase (to simplify the dataset used to train our classifier, Section 4).

[4] As seen in Figure 3, this response type comprised three options: REPHRASE OBJECT, REPHRASE POSITION and REPHRASE ALL. But we merged them into just REPHRASE owing to their low frequency in the dataset (Table 1).

	ResponseCorpus		*AuthorCorpus*	
Response type	Low risk	High risk	Low risk	High risk
Do	61.3%	45.5%	56.2%	50.3%
Confirm	8.9%	17.8%	14.4%	20.2%
Choose	20.2%	23.3%	22.9%	22.9%
Rephrase	9.6%	13.4%	6.5%	6.5%

Table 1: Response type distribution under high- and low-risk conditions

requests. The distribution of their response types appears in Columns 4 and 5 of Table 1.

Stage 2 – Rating responses to the same requests After 1.5-2 years, we were able to reach 34 participants from Stage 1, and we built *RatingsCorpus* as follows. Each participant was shown the requests they had seen before (without alerting them to this fact) together with several candidate responses. They were then asked to rate the suitability of each response on a 1-5 Likert scale under the low- and high-risk conditions.

The candidate responses were sourced from the response types chosen by the participant (*ResponseCorpus*) and the author (*AuthorCorpus*) in Stage 1, and the response types returned by a classifier trained on *AuthorCorpus* (Section 4).[5] In addition, for every Do response from Stage 1, we also presented a Confirm response in Stage 2, and vice versa. Clearly, if more than one source had the same response type for a request, this response type was presented only once in Stage 2. Figure 4 in Appendix A displays a screenshot of Stage 2 survey questions regarding the same request as that in Figure 3, presented to the same participant.

Two Stage 2 responses, viz Do and Rephrase, are direct renditions of the corresponding Stage 1 response types. However, to enable participants to rate Confirm and Choose response types, we needed to refer to specific objects. We decided to use images to mimic pointing in Confirm responses (e.g., "Do you want this [PICTURE]?") and in Choose responses with two or three candidate objects (e.g., "There are two things on the table, do you want this [PICTURE 1] or that [PICTURE 2]?"). We restricted the number of Choose responses with images because we deemed it unnatural to point to more than three things.[6] In addition, all Choose responses were realized as text only, e.g., "There are two things on the table, which one do you want?". That is, there were two Choose responses with two or three candidate objects, and one Choose response with more candidate objects. Figure 4 illustrates two Choose responses, a Confirm response and a Do response.

1 Is there an ASR output with all correct words?
2 % of wrong words in the top ASR output
3 % of wrong words in all ASR outputs
4 % of ASR outputs with all correct words

Table 2: Features that reflect the ASR's confidence

4 Using a Classifier to Select Responses

One of the aims of this project is to determine whether we can generate acceptable responses using a classifier trained on a small non-target but relevant corpus. As noted in Section 3, in order to simplify the classifier, we removed descriptions with more than one prepositional phrase. Hence, most descriptions have semantic segments corresponding to an object, a position specifier and a landmark (only 22 (7.5%) descriptions have no prepositional phrase, e.g., "the big pink ball").

4.1 Classification features

To extract features of interest, we assume an SLU system that returns several ranked interpretations, and can represent (a) the ASR's confidence in the correctness of its candidate outputs, and (b) how well an interpretation (in the context of the room) matches a given description.

We employed the output of the SLU system described in (Zukerman et al., 2015), and for each description, we automatically extracted features that represent the above two types of information. We also included information about situational risk (high or low); and for *ResponseCorpus*, we added the participants' demographic characteristics gender, English nativeness, age and education, and the difference between their risk-proneness and risk-aversion scores (Section 3).

Features that reflect the ASR's confidence. These features are shown in Table 2. They reflect the ASR's "opinion" of the correctness of its output, rather than the ground truth. The last feature is noteworthy because the ASR may have high confidence in a few ASR outputs, e.g., "the *flower* on

[5] We chose this classifier as it posts high accuracy when trained with limited data, while at the same time, representing a "worst case" for *ResponseCorpus*, as it was trained on a different corpus (the difference between the corpora is statistically significant, χ^2 with *p-value* < 0.05).

[6] Only 10 (7.9%) Choose responses under both risk conditions had more than three candidate objects.

1 # of interpretations with similar total match score to that of the top-ranked interpretation	(×1)
2 How well the relative position of OBJECT and LANDMARK in an interpretation matches the position specified in the description	(×10)
3 Lexical-match score of the OBJECT, LANDMARK and POSITION SPECIFIER in an interpretation with the corresponding semantic segment in the description	(×30)
4-6 Other match scores of each OBJECT and LANDMARK in an interpretation with the corresponding semantic segment in the description	
4 Colour match score	(×20)
5 Size match score	(×20)
6 # of *Unknown* modifiers	(×20)

Table 3: Features extracted from top-10 SLU system interpretations

the table" and "the *flour* on the table", even if only one is intended by the speaker.

Features that represent how well an interpretation matches a description. These features are summarized in Table 3. They are calculated for the top-N interpretations returned by the SLU system, where $N = 10$ (in this system, the correct interpretation is among the top ten in about 90% of the cases). The scores calculated by the SLU system for these features are combined into a total match score for each interpretation, which determines its ranking. For instance, given the description "the brown stool near the table", two stools in Figure 1(d) have a high total match score, as both are brown and near the table: the stool to the right of the table and stool **L**, which is to the left of the table. However, since the former stool is closer to the table, it has a slightly higher total score, and is ranked first, while stool **L** is ranked second.

The first feature in Table 3 represents the ambiguity of a description through the similarity between the total match score of the top-ranked interpretation and that of subsequent interpretations. We encode this similarity as the ratio between the total score of the i-th interpretation ($i = 1, \ldots, N$) and the total score of the top-ranked interpretation. All the interpretations whose ratio is above an empirically-derived threshold are deemed similar to the top-ranked interpretation.

The second feature, computed for each of the top-N interpretations, represents the goodness of the match between the position of the OBJECT in the interpretation (i.e., in the room) and its requested position in the description. For example, both stools in Figure 1(d) are *near* the table, but the position match score of the stool to the right of the table is higher than that of stool **L**.

The rest of Table 3 contains features that represent the quality of the match between individual elements in an interpretation and their corresponding semantic segments in the given description. Feature #3 represents how well the canonical name of each element in an interpretation matches the corresponding lexical item in the description. For instance, the terms "stool" and "table" respectively match perfectly the terms that designate stool **L** and the yellow table in Figure 1(d). However, if the speaker had said "ottoman", the lexical match with the canonical term for stool **L** would have been poorer.

Features #4-6 pertain to intrinsic attributes of things, which are normally stated as noun modifiers in a description. They are computed for the OBJECT and LANDMARK of each of the top-N interpretations. Following Zukerman et al. (2015), we have focused on colour and size modifiers, designating other modifiers, e.g., composition or shape, as *Unknown*. Features #4 and #5 respectively reflect the goodness of a match between the color and size of an OBJECT or LANDMARK in an interpretation and the colour and size specifications in the corresponding semantic segment in the given description. For example, a request for a "brown stool" in the context of Figure 1(d) returns a high colour match with stool **L**, while a request for a "blue stool" would return a low colour match. Finally, the match score for Feature #6, which pertains to *Unknowns*, e.g., "the *plastic* stool", reflects the badness of a match.

4.2 Classifying responses

We considered several classification algorithms to learn response types from the corpora collected in Stage 1 of our experiment (Section 3):[7] Naïve Bayes, Support Vector Machines, Decision Trees, Random Forest (RF) and Recurrent Neural Nets

[7] We tried over- and under-sampling to deal with the large majority class (DO, Table 1), and applied Principal Components Analysis to reduce the number of features, but these measures did not affect classifier performance.

Response type	*ResponseCorpus + Gender & English + RiskPronenessDiff*		*AuthorCorpus*	
	Precision	Recall	Precision	Recall
Do	0.77	0.83	0.945	0.945
Confirm	0.44	0.41	0.842	0.842
Choose	0.77	0.72	0.985	0.985
Rephrase	0.70	0.55	1.00	1.00
Accuracy	**0.72**		**0.94**	

Table 4: Per-class and overall classifier performance

(RNNs). RF yielded the best performance for both *ResponseCorpus* and *AuthorCorpus* (RNNs under-performed, as there were not enough data).

Table 4 displays the per-class and overall performance of the RF classifier with 10-fold cross validation for both corpora. As seen in Table 4, RF performed much better for *AuthorCorpus* than for *ResponseCorpus*. This is attributable to the consistency of the 584 ratings provided by one person in *AuthorCorpus*, compared to the variability among participants in *ResponseCorpus* (different participants selected different responses for requests that had the same features).

The demographic features gender and English nativeness and the difference between riskproneness and risk-aversion scores mitigated the impact of speaker diversity in *ResponseCorpus* (age and education had no effect). In addition, situational risk had some influence on classification results in *ResponseCorpus*. This is consistent with the observation that the vast majority of the differences between the low- and high-risk condition were due to changes from Do to more conservative response types, in particular Confirm (represented in Columns 2 and 3 in Table 1). Despite this, most of the misclassifications were also between Do and Confirm.

Although the performance of the RF classifier on *ResponseCorpus* is disappointing, this result is tangential to the main thrust of this paper. In Section 5, we examine participants' attitudes toward responses obtained from the RF classifier trained on *AuthorCorpus* (which is significantly different from *ResponseCorpus*, Section 3).

5 Results

The main objective of our experiment is to determine whether participants' attitudes toward responses remain consistent over time. That is, how well do participants like their own previous responses? And do they prefer them to other responses? As mentioned in Section 3, these other responses were sourced from the response types in *AuthorCorpus* and the response types chosen by the RF classifier trained on *AuthorCorpus*.

In addition, we sought to gain insights about the feasibility of using a classifier trained on the responses of one person, and to determine the influence of situational risk on people's attitudes toward response types.

Hypotheses pertaining to fewer than 200 samples were tested using Wilcoxon matched-pairs signed-rank test, and for more than 200 samples, we used the Normal approximation of this test (Siegel and Castellan, 1988).

How well do people like their previously selected response types? In order to answer this question, we had to address the following issues:

1. In Stage 1, participants selected a response type for each request, while in Stage 2, they rated responses. To compare Stage 1 selections to Stage 2 ratings, we ascribed ratings to the response types selected in Stage 1. In order to account for participants' rating bias, we assigned to each response type selected by a participant in Stage 1 the highest rating this participant gave to any response in Stage 2 (87% of these highest ratings were 5 – the maximum on the Likert Scale, Section 3).
2. In Stage 2, we offered two options for Choose response types with two or three candidate objects: Choose+pictures and Choose+text (Section 3). For each description, we assigned to a Stage 2 Choose response type the maximum of the ratings of the two options.

We tested the hypothesis that participants' Stage 1 response types yield highly rated responses in Stage 2 under both risk conditions. The result of this test was that *participants' Stage 2 ratings of responses sourced from their own Stage 1 response types were* ***significantly lower*** *than the ratings ascribed to these Stage 1 response types under the low- and high-risk conditions* (*p-value* $\ll 0.01$).

Figure 2 displays a histogram of the differences between the ratings ascribed to Stage 1 response types and the ratings given to the corresponding responses in Stage 2 under both risk conditions. For example, the leftmost bars indicate that the ratings of 159 response types under the low-risk condition and 123 response types under the highrisk condition did not change between Stage 1 and

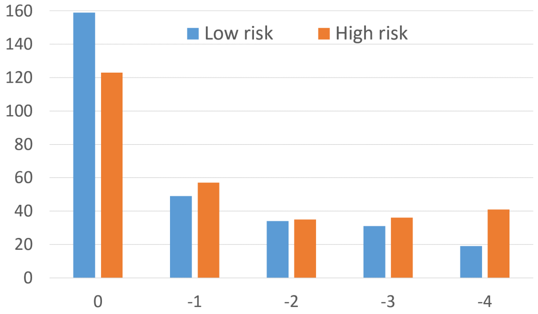


Figure 2: Differences between ratings ascribed to Stage 1 response types and ratings of the corresponding Stage 2 responses under low- and high-risk conditions

Stage 2 (the difference is 0). In other words, participants lowered their ratings of 133 response types under the low-risk condition and 169 response types under the high-risk condition. Do (majority class) accounts for 71% of these downrated response types under the low-risk condition, and 60% under the high-risk condition.

Do users prefer their previously selected response types to other response types? To answer this question, for each risk condition, we collected the participants' Stage 1 response types that *differ* from those in *AuthorCorpus* for the same request, and their response types that differ from those chosen by the RF classifier trained on *AuthorCorpus*.

Table 5 compares participants' ratings of responses (R_{S1}) sourced from their Stage 1 response types ($S1$) with their ratings of responses (R_d) sourced from *different* response types (d) selected by the RF classifier for the same requests under the low- and high-risk conditions. In total, 107 response types chosen by the classifier differ from the participants' selected response types under the low-risk condition, and 126 under the high-risk condition. In 47 of the low-risk cases and 46 of the high-risk cases, the responses sourced from the classifier's response types received a higher rating than the responses sourced from the participants' own response types (the results are similar for *AuthorCorpus*). Table 6 illustrates two of these low-risk cases, and two of these high-risk cases. For instance, in the high-risk example pertaining to Figure 1(a), the participant chose Rephrase in Stage 1, but gave it a rating of 1 in Stage 2, while Confirm received a rating of 5.

As seen in Table 5, under the low-risk condition, participants generally preferred the responses sourced from the classifier response types, while the opposite effect was observed under the high-risk condition (these findings are corroborated by the results in Table 7). Nonetheless, when we tested the hypothesis that participants liked responses sourced from their own previous response types as much as responses sourced from different response types in *AuthorCorpus* and different response types chosen by the classifier, both tests returned the same result: *there were **no statistically significant differences** between users' ratings of responses sourced from their own Stage 1 response types and their ratings of responses sourced from different response types under the low- and high-risk conditions* (*p-value* > 0.15).

Users' Stage 1 response type ($S1$) versus a *different* response type (d)	Low risk	High risk
$Rating(R_{S1}) > Rating(R_d)$	32	55
$Rating(R_{S1}) = Rating(R_d)$	28	25
$Rating(R_{S1}) < Rating(R_d)$	47	46
# of requests where $S1 \neq d$	107	126

Table 5: Comparison between participants' ratings of responses sourced from their Stage 1 response types and responses sourced from different classifier-selected response types

How does situational risk affect participants' attitudes toward different response types? As seen in Table 1, the proportion of Dos in *ResponseCorpus* decreased under the high-risk condition, while the proportion of the other response types increased (the difference between the low- and high-risk response types is statistically significant, χ^2 with *p-value* $\ll 0.01$). This indicates that participants preferred more conservative (risk-averse) response types under the high-risk condition.

Figure 2 suggests that participants were also more critical of their own previous response types under the high-risk condition than under the low-risk condition (they reduced the ratings of 169 response types under the high-risk condition compared to only 133 under the low-risk condition). This observation is confirmed by the mean ratings of the Stage 2 responses in our corpora under the low- and high-risk conditions, which are shown in Table 7 for the responses sourced from *ResponseCorpus* and the responses obtained from the RF classifier (the *AuthorCorpus* results are similar).

In addition, the ratings of Do and of both versions of Choose were significantly lower under the high-risk condition than under the low-risk condition (*p-value* $\ll 0.01$ for Do and Choose+text, and *p-value* < 0.05 for Choose+pictures). In con-

Top four ASR outputs	a. get the paint on the wall b. get the paint on the walls c. get the paint on the world d. *get the painting on the wall*	a. get the green light next to the blue plate b. get the green light next to the Blue Plate c. get the green light next to the blue planet d. get the green light next to the blue plates
Figure, requested object	1(a), **C**	1(b), **E**
Situational risk	High	High
Stage 1 response type	Rephrase (rating: 1)	Choose (rating: 1)
Stage 2 preferred response type	Confirm (rating: 5)	Confirm (rating: 5)

Top four ASR outputs	a. *move the green book rack* b. move the Greene book rack c. move the Green Book rack d. move the green book RAC	a. get the blue light on the left corner of the table b. *get the blue plate on the left corner of the table* c. get the bloop light on the left corner of the table d. get the Blue Planet on the left corner of the table
Figure, requested object	1(d), **K**	1(c), **H**
Situational risk	Low	Low
Stage 1 response type	Do (rating: 1)	Choose (rating: 3)
Stage 2 preferred response type	Confirm (rating: 4)	Do (rating: 5)

Table 6: Examples where users gave lower ratings in Stage 2 to responses sourced from their selected Stage 1 response types than to responses sourced from different response types chosen by the RF classifier; the correct ASR output is italicized

ResponseCorpus		RF Classifier	
Low risk	High risk	Low risk	High risk
3.99 (1.31)	3.59 (1.49)	4.09 (1.29)	3.54 (1.49)

Table 7: Mean (Stdev) of response ratings under low- and high-risk conditions

trast, no statistically significant differences were found with respect to Confirm and Rephrase under the two risk conditions. Also, participants preferred Confirm to Do and Choose+pictures to Choose+text under both risk conditions (*p-value* $\ll 0.01$).

These findings suggest that situational risk influences the acceptability of certain response types, but further research is required to identify these response types in a broader context.

6 Conclusion

We have offered a longitudinal study where participants initially selected response types for ASR outputs of spoken requests; and after some time, they rated responses sourced from their own response types, as well as responses sourced from other response types. Our results show that the participants did not think that their original choices were the best, and that overall, they had the same opinion of responses sourced from their own response types, the response types chosen by one of the authors and those selected by a classifier trained on the response types of the author. These findings suggest that, at least in the context of one-shot dialogues with a household robot, people's response preferences at a particular point in time may not reflect their general attitudes, and that various reasonable responses may be equally acceptable. Our results also indicate that, at least in this context, a classifier trained on a small non-target but relevant corpus may yield adequate responses.

Our experiment also distinguished between two types of situational risk: low and high. We found that risk influences people's general attitudes toward responses — they were more risk averse and critical under high-risk conditions than under low-risk conditions. However, this attitude was directed toward some response types (Do and Choose) and not others (Confirm and Rephrase). This finding, if generalized, may influence response type selection.

The implications of our findings for deep-learning systems are that training on a single best response may be unjustified, as several responses are equally acceptable. Further studies are required to determine whether our findings generalize to longer dialogues in more complex domains. If this is the case, (PO)MDP/RL systems do not need to take into account people's preferences when generating a response. However, if extra-linguistic factors such as risk come into play, they should be incorporated into policy-learning algorithms to bias response selection in favour of risk-sensitive responses preferred by people. Finally, our findings regarding rating inconsistency over time may affect the results of comparative studies, such as that of Liu et al. (2016).

7 Acknowledgments

This research was supported in part by grants DP1201001030 and LP150100060 from the Australian Research Council.

References

X. Amatriain, J.M. Pujol, and N. Oliver. 2009. I like it ... I like it not: Evaluating user ratings noise in recommender systems. In *UMAP'2009 – Proceedings of the 2009 Conference on User Modeling Adaptation and Personalization*, pages 247–258, Trento, Italy.

L. Carlson. 1983. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel Publishing Company, Dordrecht, Holland, Boston.

B. Dhingra, L. Li, X. Li, J. Gao, Y.N. Chen, F. Ahmed, and L. Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *ACL'17 – Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 484–495, Vancouver, Canada.

K. Funakoshi, M. Nakano, T. Tokunaga, and R. Iida. 2012. A unified probabilistic approach to referring expressions. In *SIGDIAL'2012 – Proceedings of the 13th SIGdial Meeting on Discourse and Dialogue*, pages 237–246, Seoul, South Korea.

M. Gašić and S.J. Young. 2014. Gaussian processes for POMDP-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 22(1):28–40.

E. Horvitz, C. Kadie, T. Paek, and D. Hovel. 2003. Models of attention in computing and communication: From principles to applications. *Communications of the ACM*, 46(3):52–57.

F. Jurčíček, S. Keizer, M. Gašić, F. Mairesse, B. Thomson, K. Yu, and S.J. Young. 2011. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *Proceedings of Interspeech 2011*, pages 3061–3064, Florence, Italy.

O. Lemon. 2011. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech and Language*, 25(2):210–221.

J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. 2016. Deep reinforcement learning for dialogue generation. In *EMNLP2016 – Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas.

W. Liao, W. Zhang, Z. Zhu, Q. Ji, and W.D. Gray. 2006. Toward a decision-theoretic framework for affect recognition and user assistance. *International Journal of Human-Computer Studies*, 64:847–873.

C-W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP2016 – Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas.

R. Moratz and T. Tenbrink. 2006. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition & Computation: An Interdisciplinary Journal*, 6(1):63–107.

N. Mrkšic, Ó.S. Diarmuid, T.H. Wen, B. Thomson, and S.J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1777–1788, Vancouver, Canada.

T. Paek and E. Horvitz. 2000. Conversation as action under uncertainty. In *UAI-2000 – Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 455–464, Stanford, California.

A. Prakash, C. Brockett, and P. Agrawal. 2016. Emulating human conversations using convolutional neural network-based IR. In *Proceedings of the Neu-IR'16 SIGIR Workshop on Neural Information Retrieval*, Pisa, Italy.

B. Rohrmann. 2005. Risk attitude scales: Concepts, questionnaires, utilizations. Technical report, University of Melbourne.

A. Said and A. Bellogín. 2018. Coherence and inconsistencies in rating behavior: Estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction*, 28:97–125.

I.V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI-17 – Proceedings of the 31st National Conference on Artificial Intelligence*, pages 3288–3294, San Fransisco, California.

S. Siegel and N.J. Castellan. 1988. *Non-Parametric Statistics for the Behavioral Sciences*, second edition. McGraw–Hill, Inc.

S. Singh, D. Litman, M. Kearns, and M. Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Artificial Intelligence Research*, 16:105–133.

K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura. 2009. Bayesian learning of confidence measure function for generation of utterances and motions in object manipulation dialogue task. In *Proceedings of Interspeech 2009*, pages 2483–2486, Brighton, United Kingdom.

B. Thomson, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, K. Yu, and S. Young. 2008. User study of the Bayesian update of dialogue state approach to dialogue management. In *Proceedings of Interspeech 2008*, pages 483–486, Brisbane, Australia.

J.G. Trafton, N.L. Cassimatis, M.D. Bugajska, D.P. Brock, F.E. Mintz, and A.C. Schultz. 2005. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, 35(4):460–470.

B-H. Tseng, F. Kreyssig, P. Budzianowski, I. Casanueva, Y-C. Wu, S. Ultes, and M. Gašić. 2018. Variational cross-domain natural language generation for spoken dialogue systems. In *SIGDIAL'2018 – Proceedings of the 19th SIGdial Meeting on Discourse and Dialogue*, pages 338–343, Melbourne, Australia.

S. Ultes, P. Budzianowski, I. Casanueva, L.M. Rojas Barahona, B-H. Tseng, Y-C. Wu, S. Young, and M. Gašić. 2018. Addressing objects and their relations: The Conversational Entity Dialogue Model. In *SIGDIAL'2018 – Proceedings of the 19th SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Melbourne, Australia.

T.H. Wen, M. Gašić, N. Mrkšic, P. Hao Su, D. Vandyke, and S.J. Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *EMNLP2015 – Proceedings ot the Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal.

J.D. Williams and S. Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

J.D. Williams and G. Zweig. 2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*.

X. Yang, Y.N. Chen, D. Hakkani-Tür, P. Gao, and L. Deng. 2017. End-to-end joint learning of natural language understanding and dialogue manager. In *ICASSP'2017 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5690–5694, New Orleans, Louisiana.

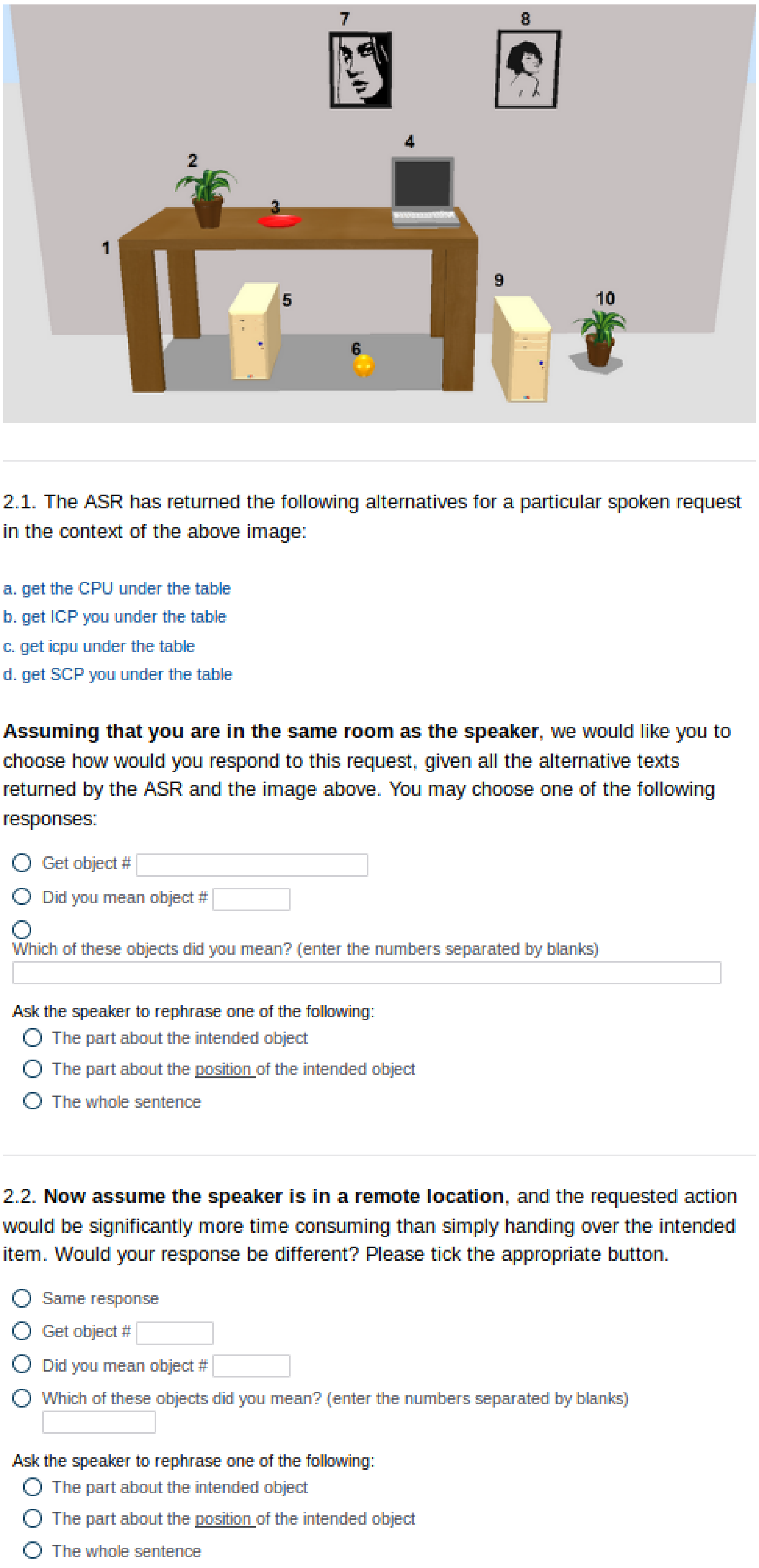
S.J. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. 2010. The Hidden Information State Model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174.

S.J. Young, M. Gašić, B. Thomson, and J. Williams. 2013. POMDP-based statistical spoken dialogue systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

T. Zhao and M. Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *SIGDIAL'2016 – Proceedings of the 17th SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Los Angeles, California.

I. Zukerman, S.N. Kim, Th. Kleinbauer, and M. Moshtaghi. 2015. Employing distance-based semantics to interpret spoken referring expressions. *Computer Speech and Language*, 34:154–185.

A Screenshots for Stage 1 and Stage 2



2.1. The ASR has returned the following alternatives for a particular spoken request in the context of the above image:

a. get the CPU under the table
b. get ICP you under the table
c. get icpu under the table
d. get SCP you under the table

Assuming that you are in the same room as the speaker, we would like you to choose how would you respond to this request, given all the alternative texts returned by the ASR and the image above. You may choose one of the following responses:

○ Get object #

○ Did you mean object #

○
Which of these objects did you mean? (enter the numbers separated by blanks)

Ask the speaker to rephrase one of the following:

○ The part about the intended object

○ The part about the position of the intended object

○ The whole sentence

2.2. **Now assume the speaker is in a remote location**, and the requested action would be significantly more time consuming than simply handing over the intended item. Would your response be different? Please tick the appropriate button.

○ Same response

○ Get object #

○ Did you mean object #

○ Which of these objects did you mean? (enter the numbers separated by blanks)

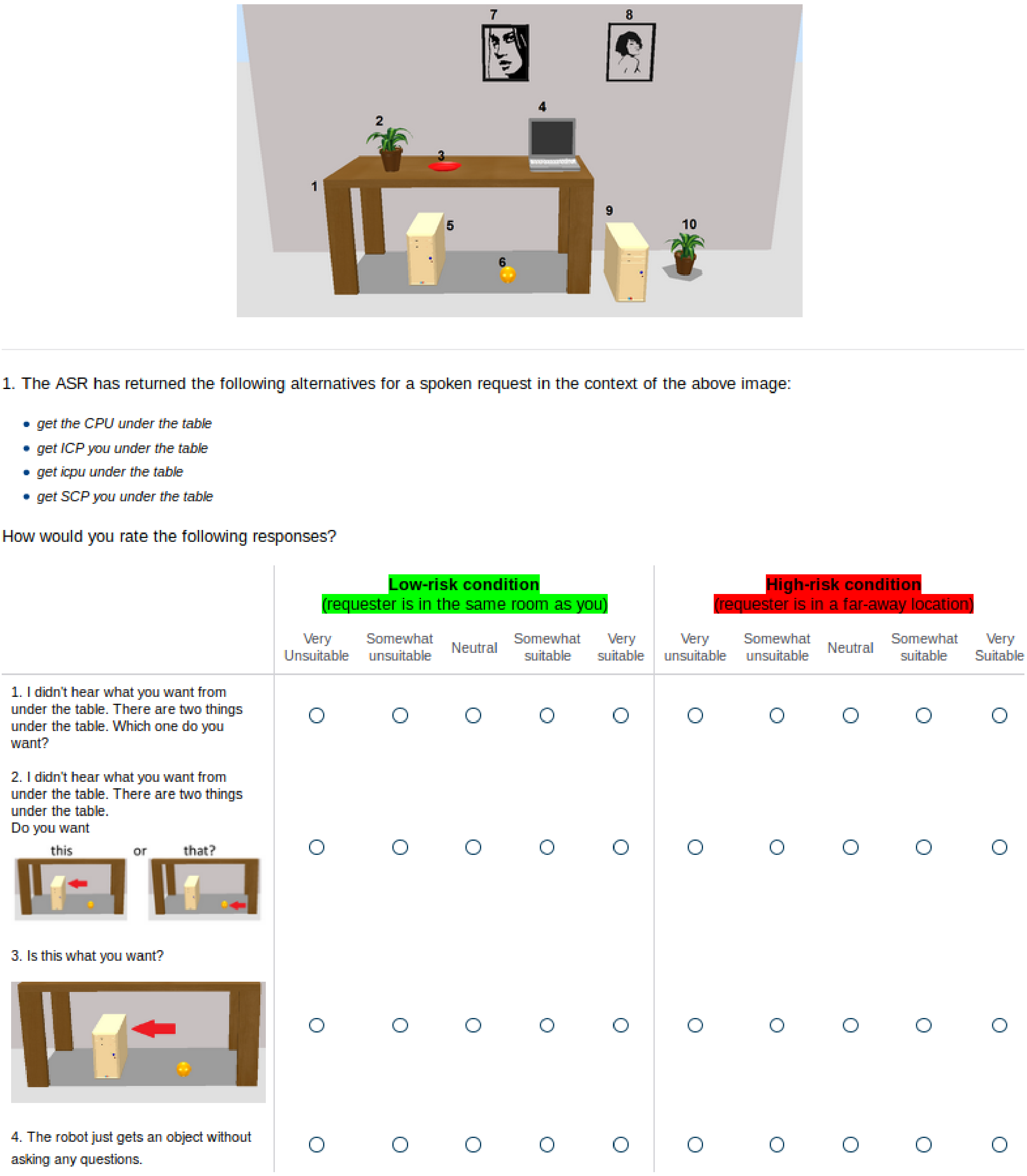
Ask the speaker to rephrase one of the following:

○ The part about the intended object

○ The part about the position of the intended object

○ The whole sentence

Figure 3: Screenshot for Stage 1



1. The ASR has returned the following alternatives for a spoken request in the context of the above image:

- *get the CPU under the table*
- *get ICP you under the table*
- *get icpu under the table*
- *get SCP you under the table*

How would you rate the following responses?

	Low-risk condition (requester is in the same room as you)					High-risk condition (requester is in a far-away location)				
	Very Unsuitable	Somewhat unsuitable	Neutral	Somewhat suitable	Very suitable	Very unsuitable	Somewhat unsuitable	Neutral	Somewhat suitable	Very Suitable
1. I didn't hear what you want from under the table. There are two things under the table. Which one do you want?	○	○	○	○	○	○	○	○	○	○
2. I didn't hear what you want from under the table. There are two things under the table. Do you want this or that?	○	○	○	○	○	○	○	○	○	○
3. Is this what you want?	○	○	○	○	○	○	○	○	○	○
4. The robot just gets an object without asking any questions.	○	○	○	○	○	○	○	○	○	○

Figure 4: Screenshot for Stage 2