Toward Automatically Measuring Learner Ability from Human-Machine Dialog Interactions using Novel Psychometric Models

Vikram Ramanarayanan
Educational Testing Service R&D
90 New Montgomery Street, #1500
San Francisco, CA
`vramanarayanan@ets.org`

Michelle LaMar
Educational Testing Service R&D
90 New Montgomery Street, #1500
San Francisco, CA
`mlamar@ets.org`

Abstract

While dialog systems have been widely deployed for computer-assisted language learning (CALL) and formative assessment systems in recent years, relatively limited work has been done with respect to the psychometrics and validity of these technologies in evaluating and providing feedback regarding student learning and conversational ability. This paper formulates a Markov decision process based measurement model, and applies it to text chat data collected from crowdsourced native and non-native English language speakers interacting with an automated dialog agent. We investigate how well the model measures speaker conversational ability, and find that it effectively captures the differences in how native and non-native speakers of English accomplish the dialog task. Such models could have important implications for CALL systems of the future that effectively combine dialog management with measurement of learner conversational ability in real-time.

1 Introduction

Advances in multimodal dialog technologies have helped improve the state of the art in interactive computer-assisted language learning (CALL) and educational assessment applications in recent years. However, while much progress has been made with respect to the technology infrastructure and automated processing required in such dialog applications, relatively less work has carefully investigated the efficacy and validity of such assessment instruments, for instance, how well they measure students' capabilities. In other words, there is relatively little investigation into the psychometrics of such CALL applications and dialog-based assessments[1].

Interactive tasks such as multi-turn conversations have had limited use as standardized assessments due in part to the difficulty of evaluating these performances. When such assessment tasks are used, the conversational performance is scored primarily using human raters (take for instance, the IELTS exam[2]). Machine scoring of complex task performances has made substantial progress, especially is the domain of written essays (Shermis, 2014), but has been limited by path complexity in interactive performances such as dialog (Graesser et al., 2005).

While technical language use, (e.g. grammar or pronunciation) might be scorable at the word or phrase grain size, pragmatic conversational ability can only be judged in the context of the conversation history, personal goals, and interpersonal dynamics. In a conversational task, for example, the "correctness" of single utterances cannot be scored independently as their function, and therefore their value, depends upon the current state of the dialog. An utterance at one stage of the conversation might be of high value while the same utterance at a different point would be detrimental. Each utterance must be evaluated based on the speaker's conversational goals, what they have already accomplished in the conversation, and what sequence of interactions might bring them closer to their goal.

Such data is unsuitable to model with traditional psychometric models that assume conditionally independent performance data, such as either classical test theory or item response theory (De Boeck and Wilson, 2004), requiring a more structured and dynamic model (Mislevy et al., 2002). It is this modeling gap that we attempt to bridge in this paper using Markov Decision Process (or MDP)-based measurement modeling (LaMar, 2018). To our knowledge, this is the first attempt at developing a psychometric model for dialog data that explicitly accounts for temporal dependencies in the observed data stream.

---

[1] Psychometrics is the field of study concerned with the theory and technique of psychological measurement, which includes the measurement of knowledge, abilities, attitudes, and personality traits. Psychometricians use a specialized set of statistical tools to create scientifically valid "standardized" assessments of various behaviors. Typically, a test is considered to have been standardized if data have been collected on large numbers of subjects using a set of structured rules for administration and scoring. These data are used to determine the mean score and the standard deviation, which the psychometrician then uses to benchmark the performance of those being tested. For more details, see Association et al. (1999) or Weiss and Zureich (2008).

[2] `https://www.ielts.org/`

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 117–126
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

While the field does need more research into psychometrics and validity of dialog-based summative assessments, there has been substantial work by the learning and formative assessment community in examining learning gains/progressions and modeling cognitive strategies in conversational tutoring applications (see for example Person et al., 2001; VanLehn et al., 2002; Heffernan and Koedinger, 2002; Michael et al., 2003; Pon-Barry et al., 2006; Rus et al., 2013). Researchers have also examined how one can perform adaptive dialog management to personalize the instruction to individual participants over the course of the interaction (Forbes-Riley and Litman, 2011; Vail and Boyer, 2014). This includes using learning progressions, natural language processing and affective computing to adaptively selecting appropriate tasks for the learner to work on, but also adapting the scaffolding while the learner is working on a tasks (Rus et al., 2013).

Such research has important implications for dialog system design as well. Particularly for CALL applications, it is important to integrate formative assessment of student ability into the dialog management process, in order to better adapt instruction to student needs, both in terms of the level of instruction (obtained in real time through measurement models) as well as the content and dialog path (decided by the dialog manager). We envision that future statistical dialog systems could combine statistical dialog management achieved using Partially Observable Markov Decision Processes or POMDPs (see for instance Young, 2006; Williams and Young, 2007; Young et al., 2010) in tandem with statistical measurement (using POMDP-based models) in order to develop more effective conversational language learning applications.

Our work also directly relates to user modeling in dialog systems. While there is plenty of theoretical work on such models (see for example, Kobsa, 1990; Kass, 2012), implemented statistical versions of user models typically estimate the probability of a particular user response given a candidate system response or a interaction history thereof (e.g., Eckert et al., 1997; Levin et al., 2000; Horvitz and Paek, 2001; Pietquin, 2005; Kim et al., 2008). However, the difference in our case is that in order to serve as a measurement model of student performance, our MDP represents the cognitive model of an ideal automated interlocutor. Given a specified set of model parameters, the MDP model can generate action (or response) probabilities for every possible conversational state, depending on a learner/user-specific latent 'conversational ability' parameter which needs to be estimated for each user. Note that for the purposes of *this* paper, we will be broadly looking at conversational ability (in achieving a certain goal), and *not* necessarily technical English language proficiency.

The rest of the paper is organized as follows: Section 2 lays out the mathematical foundations of how MDP models can be used to model learner ability, including the equations for statistical parameter estimation. Section 3 then describes the dialog infrastructure used along with details regarding the conversational task and crowdsourcing data collection, followed by the formulation of the task-specific MDP for our use case in Section 4. Section 5 analyzes the results of running the model on our dataset and studies how well the model differentiates between native and non-native speakers (who are potential language learners) of English, with example dialogs included for illustration purposes. Finally, we conclude with a discussion of the current state of the art and outstanding issues for future research.

2 Markov Decision Process Measurement Models

As an extension of inverse reinforcement learning, partially observable Markov Decision Processes (POMDPs) have recently been used to represent a cognitive model that describes both human decision making and people's ability to infer the goals and beliefs of others. Baker et al. (2011) describe a "Bayesian theory of mind" in which cognition is modeled as a POMDP. They hypothesize that people act based on their beliefs, modeled by the state space, action set, and transition functions, and in accordance to their desires, which are modeled by the reward structure. With this cognitive framing, POMDPs can be used for measurement within a goal-directed task by comparing actions selected by human participants with the model's predicted probability of those actions (LaMar, 2018). The model and estimation algorithms will be described briefly below; for full details can be found in LaMar (2018). Note that in this work we utilize the more constrained MDP, in which the problem state is assumed to be observable, but extensions to full POMDP models are a natural next step.

2.1 Mathematical Formulation

As a decision model, the MDP defines the probability of selecting of an action $a \in A$ given a specific state of the task $s \in S$. This probability, $p(a|s)$, is known as the policy. Action selection occurs within the context of a reward function $r(s,a,s')$, which specifies the immediate reward for taking action a in state s and entering state s' and a transition model $p(s'|s,a)$, which is the probability of transitioning to a state s' given that action a was taken in state s. An additional parameter $\gamma \in [0,1]$, known as the discount parameter, represents the relative value of future versus immediate rewards. From this specification, one can calculate the Q function, which is the expected sum of discounted rewards obtained by taking action a while in state s,

$$\text{Q(s,a)}=\sum_{s' \in S} p(s'|s,a)\left(r(s,a,s') + \gamma \sum_{a' \in A} p(a'|s') Q(s',a')\right). \quad (1)$$

Note that $\sum_{a' \in A} p(a'|s')Q(s',a')$ is the expected value of the next state, marginalized over the possible next actions. Thus the quantity inside the large parentheses is the sum of the immediate reward and the discounted value of the future state. The expectation of this sum is then taken over all possible states s' that might result from action a in state s. The Q function is recursive, as the value of a state is defined using the Q function itself, but can be calculated using dynamic programming (Howard, 1960).

When MDPs are used in the context of artificial agents, they generally employ an optimal policy which selects the action that maximized Q in each state. To model human performance, however, optimal decision making is not assumed. Instead a Bolzmann policy is used (Baker et al., 2009),

$$p(a|s) \propto e^{\beta Q(s,a)}, \tag{2}$$

where $\beta \in [0,\infty)$ represents the decision maker's ability to choose actions that will result in higher total rewards. As β increases, the probability choosing an optimal action increases. When β goes to zero, actions are selected uniformly at random from the action set.

2.2 MDPs for Measurement and Inference

Researchers have recently extended the MDP framework to study the quality of inferences that can be made about student/learner cognition based on records of action; for instance, to model learner goals and beliefs (Rafferty et al., 2015; Baker et al., 2009), to model inquiry strategies (LaMar et al., 2017), and to model student decision making ability (LaMar, 2018). Using the Boltzmann policy (Eq. 2), the MDP model can be seen as a generative latent-trait model provided that the latent traits of interest can be formulated as parameters of the model. While elements of the reward function and the transition model can be parameterized for inference about the decision maker's goals and beliefs, here we focus on the capability parameter β_j, a *person-specific* Boltzmann parameter, indicating a person's capability to optimally solve the given problem. The formulation of the Q function remains as in Equation 1, except that we note explicitly the dependency upon the capability parameter β_j. The conditional probability of student j selecting action a when in state s now becomes

$$p(a|s,\beta_j) = \frac{\exp\left(\beta_j Q(s,a|\beta_j)\right)}{\sum_{a'\in A} \exp\left(\beta_j Q(s,a'|\beta_j)\right)}. \tag{3}$$

If the reward and transition parameters are fixed to objectively correct values, the Q function acts as a scoring function, determining the relative value of the actions available in each state. The β_j parameter is then similar to a traditional ability parameter in IRT, measuring the extent to which the highest valued action is taken at each decision point.

2.3 Parameter Estimation

The observed data for student j consist of a sequence of state-action pairs,

$$O_j = \{(s_{1j},a_{1j}),(s_{2j},a_{2j}),\ldots(s_{N_j j},a_{N_j j})\}, \tag{4}$$

where N_j is the total number of actions taken by the student. Each pair indicates a state and the action taken in that state.

The Markov property applies to this model, allowing us to take each action to be conditionally independent, conditioned upon student capability and the system state in which the action was taken. Thus the probability of the observed data can be written as

$$\begin{aligned} p(O_j|\beta_j,\mu,\sigma) &= \prod_{t=1}^{T_j} p(a_{tj}|s_{tj},\beta_j,\mu,\sigma) \qquad (5)\\ &= \prod_{t=1}^{T_j} \frac{\exp\left(Q(s_{tj},a_{tj}|\beta_j)\beta_j\right)}{\sum_{a'\in A}\exp\left(Q(s_{tj},a'|\beta_j)\beta_j\right)}. \\ \text{where } \beta_j &\sim \mathit{Lognormal}(\mu,\sigma^2) \qquad (6) \end{aligned}$$

where the optimal value of the person-specific ability parameter, $\hat{\beta}_j$, can be estimated by finding the value of β_j that maximizes this likelihood:

$$\hat{\beta}_j = \operatorname*{argmax}_{\beta_j} p(O_j|\beta_j,\mu,\sigma) \tag{7}$$

To estimate the population parameters of the lognormal distribution[3], μ and σ, we use marginal maximum likelihood (MML), marginalizing over the person-specific parameter distributions. The person-specific β_j can be estimated either using maximum a-posteriori (MAP) or maximum likelihood estimation (MLE) methods. With smaller population sizes the MLE estimation has been found to be more robust and is used for this study. Both the MML and MLE estimations are performed using a two-phase numerical optimization with a global optimization algorithm followed by a local optimization algorithm, both drawn from the *nlopt* library. Gaussian quadrature is used for the approximation of the integrals and the Q-function is approximated using policy iteration methods.

3 Dialog Data

3.1 Dialog System

We use an open-source dialog system[4] to develop a text-based chatbot application. But note that this work is not limited to or dependent on the dialog system being used. Indeed, there are multiple academic (Olympus (Bohus et al., 2007), Alex (Jurčíček et al., 2014), Virtual Human Toolkit (Hartholt et al., 2013), OpenDial[5], etc.) and industrial (Voxeo[6], Alexa [7], etc.) implementations of dialog systems, any of which can be

---

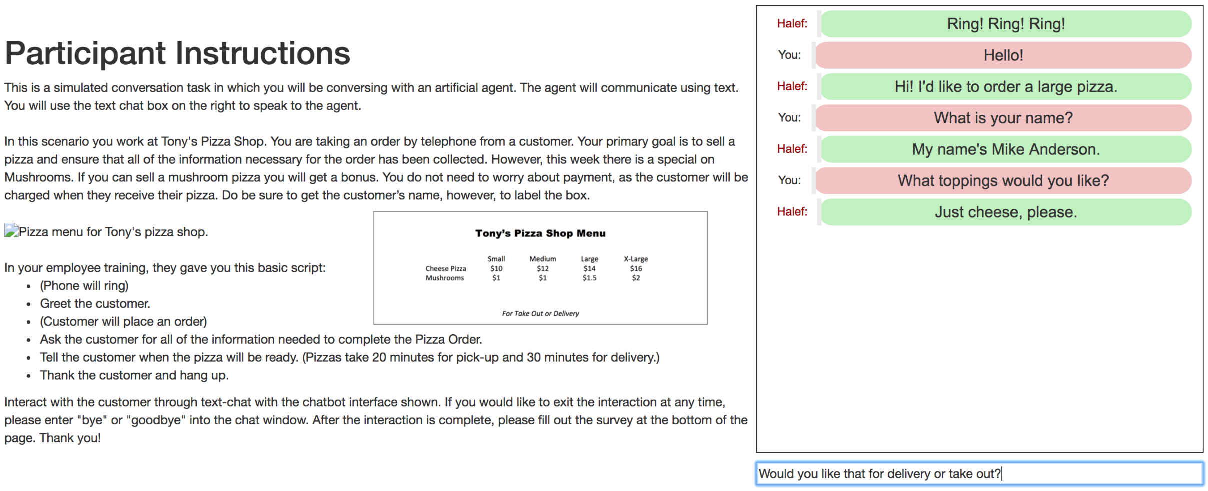
[3] We chose to model β_j using the Lognormal distribution as it has a non-negative range and has been used to describe growth patterns in nature. Moreover, person-specific "ability" can be argued to be the result of a growth process (multiplicative rather than additive).

[4] `Link to software anonymized.`

[5] `http://www.opendial-toolkit.net`

[6] `https://voxeo.com/prophecy/`

[7] `https://developer.amazon.com/alexa`



Participant Instructions

This is a simulated conversation task in which you will be conversing with an artificial agent. The agent will communicate using text. You will use the text chat box on the right to speak to the agent.

In this scenario you work at Tony's Pizza Shop. You are taking an order by telephone from a customer. Your primary goal is to sell a pizza and ensure that all of the information necessary for the order has been collected. However, this week there is a special on Mushrooms. If you can sell a mushroom pizza you will get a bonus. You do not need to worry about payment, as the customer will be charged when they receive their pizza. Do be sure to get the customer's name, however, to label the box.

Pizza menu for Tony's pizza shop.

Tony's Pizza Shop Menu

	Small	Medium	Large	X-Large
Cheese Pizza	$10	$12	$14	$16
Mushrooms	$1	$1	$1.5	$2

For Take Out or Delivery

In your employee training, they gave you this basic script:

- (Phone will ring)
- Greet the customer.
- (Customer will place an order)
- Ask the customer for all of the information needed to complete the Pizza Order.
- Tell the customer when the pizza will be ready. (Pizzas take 20 minutes for pick-up and 30 minutes for delivery.)
- Thank the customer and hang up.

Interact with the customer through text-chat with the chatbot interface shown. If you would like to exit the interaction at any time, please enter "bye" or "goodbye" into the chat window. After the interaction is complete, please fill out the survey at the bottom of the page. Thank you!

Halef: Ring! Ring! Ring!
You: Hello!
Halef: Hi! I'd like to order a large pizza.
You: What is your name?
Halef: My name's Mike Anderson.
You: What toppings would you like?
Halef: Just cheese, please.

Would you like that for delivery or take out?

Figure 1: Example webpage screenshot of the text dialog interface that participants might see for the task described in this paper.

used, but many of these often use special architectures, interfaces, and languages paying relatively less attention to existing W3C and other industry standards (see Ramanarayanan et al. (2017) for more details). We however choose to use the *Anonymous* cloud-based dialog system for its standards-compliance, modularity and flexibility in developing both text- and speech-based applications. In this study we will limit ourselves to text-based dialog for simplicity.

3.2 Conversational item design

This study leverages a conversational practice task developed for English language learners, where subjects are asked to pose as a customer services representative at a pizza restaurant, and field an order from an automated customer (played by the dialog system). See Figure 1 for a screenshot of the web-based dialog interface that participants interacted with. Participants are instructed that their primary goal is to sell a pizza while ensuring that they collect all information necessary to complete the order (such as the name of the customer, his address if delivery is requested, etc.). They are further instructed that if they manage to sell the customer mushroom toppings, they will be awarded a bonus for task performance. We used regular expressions to perform the natural language understanding. Figure 2 depicts the dialog flow of the conversational item. Recall that for the purposes of *this* paper, the target of measurement is the student's ability to navigate conversational conventions and achieve the pre-specified task goal (to maximize the pizza sale) through conversation with the automated customer, and not their technical language skills.

3.3 Crowdsourcing data collection

We used Amazon Mechanical Turk for our crowdsourcing data collection experiments. Crowdsourcing has been used in the past for the assessment of dialog systems as well as for collection of dialog interactions (see for instance (McGraw et al., 2010; Rayner et al., 2011; Jurcıcek et al., 2011; Ramanarayanan et al., 2016)). In addition to interacting with the text chatbot interface to complete the conversational task, workers were requested to fill out a 2-3 minute survey regarding different aspects of the interaction, such as their overall experience, how engaged they felt while interacting with the system, how well the system understood them, and basic demographic information. Particularly relevant for this study are callers' self-reported first language, and their ratings of system performance, defined as a qualitative measure of how the system performed as per caller expectations and whether the system responses were appropriate. In all we collected and analyzed dialogs from 390 participants, 54% of which self-reported as native English language speakers and 70% of which were male, primarily in the 20–40 age range. See Tables 3–7 for example dialogs.

4 MDP Model for the Pizza Dialog Task

Table 1: The action set and transition function for the MDP PizzaOrder cognitive model.

Action	Prob	Customer ...
RequestToppings	0.6	requests cheese pizza
RequestToppings	0.4	requests mushroom pizza
SuggestMushroom	0.4	requests cheese pizza
SuggestMushroom	0.6	requests mushroom pizza
AskDelivery	0.5	wants delivery
AskDelivery	0.5	wants take-out
AskName	1.0	gives name
AskSize	1.0	orders large pizza
AskAddress	1.0	gives address
AskPhone	1.0	gives phone number
SayOther	1.0	no effect
EndConversation	1.0	hangs up

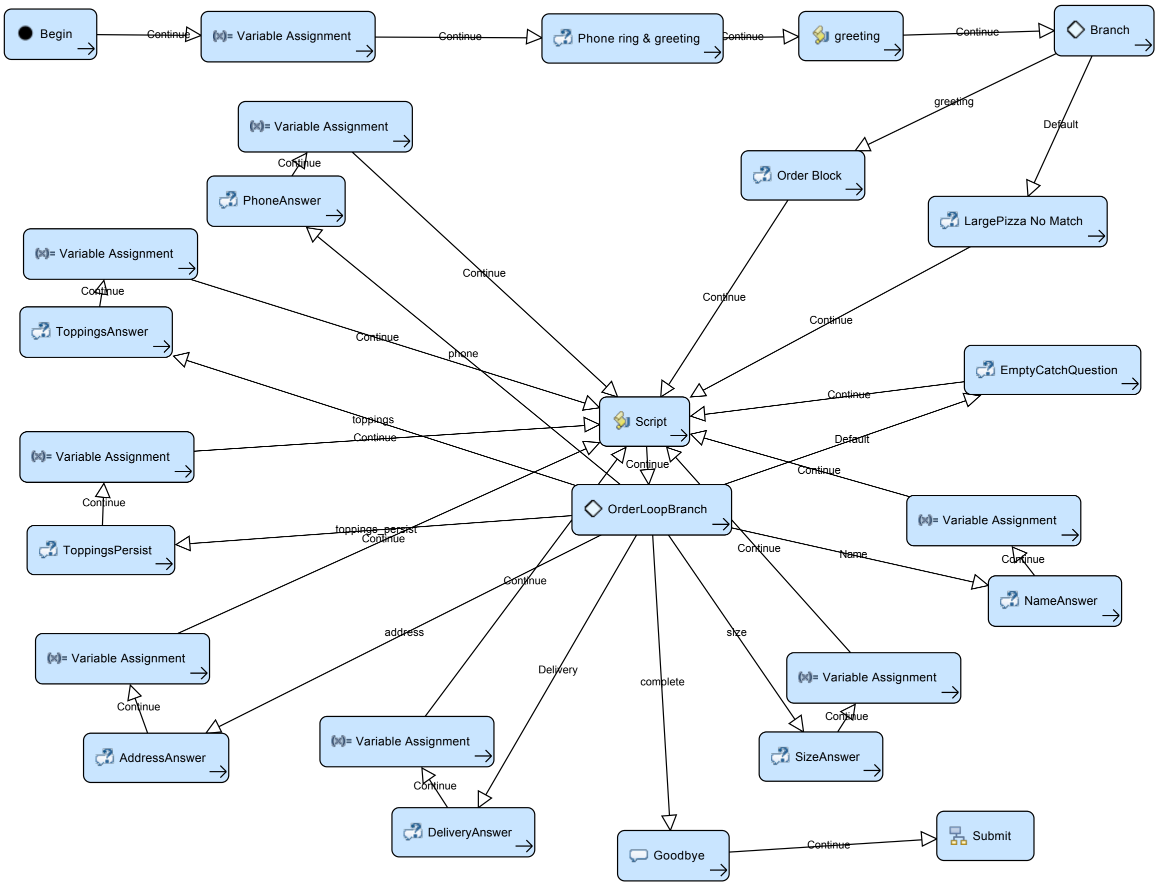


Figure 2: Workflow design of the pizza restaurant-based conversational task deployed and analyzed in this paper. The dialog flow employs a slot-filling structure, where the customer service representative has to ask the automated customer multiple questions in order to fill slots on an order form.

Table 2: The reward table for the simple Pizza Order task

Reward Type	Value	Trigger
PizzaOrdered	3	end-of-call and got all needed info
Mushrooms	1	end-of-call and PizzaOrdered + mushrooms ordered
CompleteInfo	0.5	end-of-call and also got phone #
AngryCustomers	-0.3	any action while customer is annoyed
TimeIsMoney	-0.2	any action

To serve as a measurement model for student performance, the MDP must represent the cognitive model of an ideal pizza shop representative. The full MDP cognitive model consists of a set of actions, a state space, the transition functions, and the reward structure. In Table 1 the action set is listed in the left column, while the transition function is partially illustrated by the probability of effects from each action. The state space is defined by a set of state variables which includes information slot boolean variables such as *gotSize*, *gotAddress*, and *gotCustomerName*. For order information which might affect the choice of future actions, we model the possible values along with a value for "unknown." For example, the *wantsMushroom* variable has three discrete values, 0 for unknown, -1 for "does not want mushrooms," and 1 for "wants mushrooms;" *wantsDelivery* is coded similarly. The possibility of customer annoyance (*isAnnoyed*), which was ommitted from Table 1 for clarity, adds complexity to the dialog task. The cognitive model assumes that every time the customer is asked a question that they have already answered they have a .5 probability of becoming annoyed. This means that while pestering the customer to order mushrooms might result in an mushroom pizza order, it also might result in an annoyed customer. Annoyed customers do not buy pizzas. The final *isSold* state variable gets set to 1 only if all the required information has been gathered:name, topping-preference, size, delivery-or-takeout, and address (in the case of delivery) *and* the customer is not annoyed. Note that the model is intended to reflect the thinking of a competent participant engaging with that task and thus includes stochastic transitions based on “likely” outcomes from different conversational moves. For example, in the test task the customer will always ask for delivery, however the cognitive model for the pizza salesperson gives the probability of the customer wanting delivery as .5, as the representative does not know whether the customer will want delivery until they ask.

The rewards are defined based on actions taken and values of the state variables. Table 2 shows the five reward or cost conditions used for this model. The biggest reward comes from completing the call with a pizza ordered, for a reward of 3 points. An additional point is awarded if the pizza order includes mushrooms and an additional half point is given for collecting the not-required, but nice-to-have phone number of the customer. When customers get annoyed, the whole conversation gets unpleasant causing a cost of 0.3 points (equivalently a reward of -0.3) for every action taken while the customer is annoyed. Finally, each and every action costs 0.2 points. This “time cost” is essential for motivating efficiency in action planning.

Given this specification, the PizzaOrder MDP can generate action probabilities for every possible conversational state, dependent upon the β_j parameter, which we take to be conversational ability as demonstrated by the task.

5 Analyses and Observations

Table 3: Example dialog from a native Hindi speaker ($log(\beta) = -2.75$).

Speaker	Dialog
Human	Hello, Welcome to Tony’s Pizza Shop, How may I help you today?
System	Hi! I’d like to order a large pizza.
Human	Of course. May I know your name please?
System	My name’s Mike Anderson.
Human	Sure sir, We have a special on mushroom pizza this week, would you be interested in that?
System	Just cheese, please.
Human	Is that a large or an extra large? A large cheese pizza is $14, andd an extra large is $16.
System	Just cheese, please.
Human	Large cheese pizzza.
System	Just cheese, please.
Human	the pizza will be ready in 30 minutes
System	Alright, thanks! Goodbye!

We ran the MDP measurement model over the full data set of 390 dialogs, estimating the person-specific conversational ability (which, again, is their ability to achieve the pre-specified task goal through conversation, and not their technical language skill) for each dialog as β_j. In our analyses we examine $log(\beta_j)$ which should follow a more familiar normal distribution. To evaluate the validity[8] of the resulting estimates, we analyzed them vis-a-vis their reported native language, with the expectation that native-English speakers would generally have higher conversational ability[9]. Recall that our sample included 209 native English speakers, and 181 participants who reported a non-English language as their native tongue.

Table 4: Example dialog from a native Tamil speaker ($log(\beta) = -1.26$).

Speaker	Dialog
Human	hello
System	Hi! I’d like to order a large pizza.
Human	Are you want mushroom pizza?
System	Just cheese, please.
Human	Are you want large one?
System	I’d like a large, please.
Human	It will take 30 minutes to deliver.
System	Alright, thanks! Goodbye!

Table 5: Example dialog from a native Dutch speaker ($log(\beta) = 0.258$).

Speaker	Dialog
Human	Hi, welcome to Tony’s Pizza Shop
System	Hi! I’d like to order a large pizza.
Human	Absolutely, what kind of pizza would you like? We have a special on mushroom pizza, where you can get a medium mushroom pizza for just $1
System	Just cheese, please.
Human	A large cheese pizza is nearly 10 times as expensive as a mushroom
System	Okay, I’d like mushrooms on my pizza as well.
Human	Absolutely, would you like it delivered or do you want to pick it up?
System	Delivery, please.
Human	And what’s your name and adress?
System	My name’s Mike Anderson.
Human	and your adress?
System	I’m sorry, I didn’t catch that. Can you repeat that, please?
Human	and your address?
System	23 Maple Street.
Human	Ok, that will take 30min
System	Alright, thanks! Goodbye!

We observed that participants who reported English as their first language had average $log(\beta_j)$ estimates 0.31 logits higher than players who reported a non-English language as their first language ($t = 3.00, df =$

---

[8] In other words, does this test actually measure what it sets out to measure?

[9] Note that native language is *not* the best variable to use to evaluate validity, merely something that might be a correlate thereof. We used this in lieu of better evidence for establishing validity. Future work will explicitly look into measuring better correlates of validity.

Table 6: Example dialog from a native Indonesian speaker ($log(\beta) = 1.57$).

Speaker	Dialog
Human	Hello, thanks for calling Tony's Pizza. How can I help you?
System	Hi! I'd like to order a large pizza.
Human	Would you like mushrooms on your pizza?
System	Just cheese, please.
Human	Will this be take out or delivery?
System	Delivery, please.
Human	May I have your address?
System	23 Maple Street.
Human	And your phone number?
System	My number's 5550134.
Human	And your name please.
System	My name's Mike Anderson.
Human	Great, we will send the pizza over in 30 minutes.
System	Alright, thanks! Goodbye!

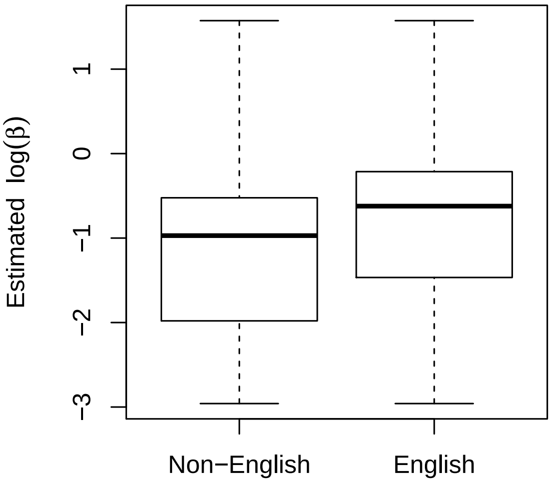


Figure 3: Estimated conversational ability by participant's self-reported primary language.

374, $p = 0.003$). Figure 3 shows a boxplot of the estimated $log(\beta)$ values grouped by native language.

We also compared the dialog-ability estimates to participant-reported satisfaction ratings with respect to the dialog system performance. Here we hypothesize that participants who used phrases that the dialog engine did not recognize would both be dissatisfied with the performance of the system and have low estimated conversational ability. In our sample, 254 participants reported that the system performed well (4 or 5 on a 5-point Likert scale), while 103 participants rated the system at a 3 or lower. Players who rated that the system performed well had an average $log(\beta_j)$ estimate 0.47 logits higher than those who rated the system poorly ($t = 3.64, df = 159.7, p < 0.001$) (Figure 4), which seems to conform with our hypothesis. However, note that these system performance ratings are subjective and might vary depending on the speaker sample and specific conversational item under study.

While these results provide, as yet, only weak validity evidence for the measurement model, they do indicate that the model is performing as expected. We also examined the actual dialogs of different participants interacting with the system in order to better understand how the model of student dialog reflects actual student performance. We have listed example dialogs of non-native participants interacting with the system of different estimated dialog ability and self-reported system performance rating. Note that these are presented as is, without correcting for errors in spelling or grammar. Table 3 shows an example dialog which was assigned a low dialog ability rating ($log(\beta_j)$) as well as a low system performance rating. In this case, while the Hindi speaker mentioned the deal on mushrooms, he asked for the pizza size again even though the automated customer had already given him that information. Per our earlier model specification, this might have 'annoyed' the automated customer. Crucially, though, he failed to ask the automated customer whether he wanted delivery or not, and subsequently his address, which resulted in a low $log(\beta_j)$ score on the task overall. Table 4 shows an example where the automated customer did not get annoyed, but it nonetheless shows clear gaps in the non-native participant's conversational competence in achieving the goal of maximizing the sale. In contrast to these examples, the Indonesian speaker (Table 6) asked the automated customer for each of the requisite pieces of information to complete the task successfully resulting in a successful interaction that received a high $log(\beta_j)$ score, despite the fact that he didn't sell the customer mushrooms. A native speaker of Dutch (Table 5) who performed well on the task in general, but was scored slightly lower ($log(\beta_j) = 0.258$) did persist in selling mushroom toppings to the automated customer while asking for his name and address, but incorrectly spelled the word 'address'. However, the participant caught this error in the next dialog turn, ultimately resulting in successful completion. Note that there were also cases that received a high $log(\beta_j)$ score with low system performance ratings, many of which were due to system natural language understanding issues. Going forward, we will aim to improve this aspect of the system to improve user experience and modeling accuracy.

6 Discussion and Outlook

We have presented a Markov decision process-based measurement model (MDP-MM) for the assessment of of learners' ability to complete a simple customer interaction dialog task. We put forth a formal mathematical description of the model including a maximum likelihood based method to estimate the parameters of the model given input data. On applying the model to crowdsourced customer services dialog interactions at a pizza restaurant, we observed that the model ability ($log(\beta_j)$) estimate is able to differentiate between native and non-native speakers of English and partic-

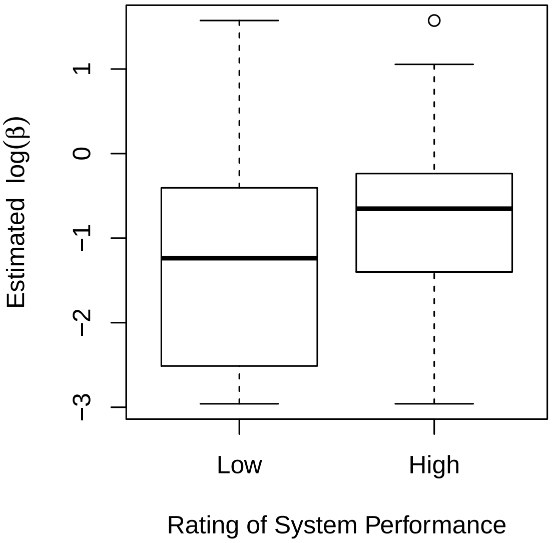


Figure 4: Estimated conversational ability by participant's rating of the system's performance, where "Low" indicates a rating of 1, 2, or 3 and "High" indicates a rating of 4 or 5 on the 5-point Likert scale.

ipant ratings of system performance in a statistically significant manner. Note that the MDP-MM is particularly useful over traditional methods of measurement when the dialogs increase in complexity and branching, and the resulting paths cannot be easily enumerated for scoring.

We plan to investigate several lines of research going forward. First, while we have shown the model's efficacy in capturing conversational ability of participants in successfully completing a given task to a certain extent, neither the degree of nativeness nor their rating of system performance are ideal correlates to establish the validity of the model. A more appropriate variable might be, for instance, an 3rd-party expert rating of their conversational ability (where experts could be English language teachers, for instance). In addition, we hand-crafted a specific set of actions, transition probabilities and rewards for the model presented in this paper based on our subjective expertise. Careful selection of these parameters is important because they directly influence model behavior. Future iterations could benefit from a more scientifically objective method of model specification. We will also need more data from more conversational items and participants to concretely establish the utility of the model and its applicability to a wide variety of dialog use cases in a statistically significant manner.

Second, while this paper has focused on conversational task ability, our longer term goal is to apply such a model to the measurement of conversational language proficiency. This will require modifications to both the task (the goals, dialog flow design, natural language understanding and dialog management logic) as well as the specific variables we measure (such as fluency, language use, vocabulary and grammatical accuracy, pragmatics and historical discourse context, among others).

Third, while the proposed model assumes that the state of the system is known at every given point of time for simplicity, relaxing this assumption is a natural next step. In such a case, we would have to use a *partially observable* extension of the MDP-MM model (or a POMDP-MM) that explicitly models the uncertainty in the observation process that estimates the state of the system at every time step.

A fourth important future research direction, as mentioned in this paper's introductory paragraphs, involves the integration of statistical measurement of student conversational ability with dialog management, especially for computer-assisted language learning (CALL) or formative assessment applications. Such integration would leverage the measurement of learner conversational ability and/or language proficiency into the dialog manager, allowing one to adapt the conversational instruction flow both based on the content of what the learner said, as well as his/her conversational ability. In addition, popular statistical dialog management modules are based on POMDPs, which might allow for easier combination with the POMDP-based measurement model into a unified model, given that both share the underlying mathematical framework. For example, in such a scenario, one could imagine that the user action model, user goal model and dialog model in a POMDP-based dialog manager (that estimate the user's next action and state, and the next dialog system state, respectively, as described in Young, 2006) would now depend (and be conditional) on the user's conversational ability and/or language proficiency estimate.

Finally, we also plan to evaluate model efficacy and integrability into a full-blown spoken dialog scenario (as opposed to text chat, as in this paper). In addition, the current paper uses simple regular expression-based natural language understanding; incorporating more accurate statistical natural language understanding modules could further improve model performance and estimation accuracy. Such improvements and the early nature of the model notwithstanding, the relative lack of previous work in measuring conversational ability in CALL dialogue and the results presented in this paper speak to the necessity and potential of such measurement models in developing more comprehensive and effective CALL applications.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. American Educational Research Association.

C.L. Baker, R. Saxe, and J.B. Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113(3):329–349.

C.L. Baker, R.R. Saxe, and J.B. Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*. page 2469–2474.

D. Bohus, A. Raux, T. Harris, M. Eskenazi, and A. Rudnicky. 2007. Olympus: An Open-Source Framework for Conversational Spoken Language Interface Research. In *Proc. of the HLT-NAACL*. Rochester, USA.

P. De Boeck and M. Wilson. 2004. *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Verlag.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, pages 80–87.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication* 53(9):1115–1136.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on* 48(4):612–618.

Arno Hartholt, David Traum, Stacy C Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now. In *Intelligent Virtual Agents*. Springer, pages 368–381.

Neil T Heffernan and Kenneth R Koedinger. 2002. An intelligent tutoring system incorporating a model of an experienced human tutor. In *International Conference on Intelligent Tutoring Systems*. Springer, pages 596–608.

Eric Horvitz and Tim Paek. 2001. Harnessing models of users goals to mediate clarification dialog in spoken language systems. In *International Conference on User Modeling*. Springer, pages 3–13.

Ronald A. Howard. 1960. *Dynamic Programming and Markov Processes*. The MIT Press, Cambridge, Mass., 1st edition.

Filip Jurčíček, Ondřej Dušek, Ondřej Plátek, and Lukáš Žilka. 2014. Alex: A statistical dialogue systems framework. In *Text, Speech and Dialogue*. Springer, pages 587–594.

Filip Jurcıcek, Simon Keizer, Milica Gašic, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2011. Real user evaluation of spoken dialogue systems using amazon mechanical turk. In *Proceedings of INTERSPEECH*. volume 11.

Robert Kass. 2012. Student modeling in intelligent tutoring systems–implications for user modeling. *User Models in Dialog Systems* page 386.

Dongho Kim, Hyeong Seop Sim, Kee-Eung Kim, J Kim, JW Sung, et al. 2008. Effects of user modeling on pomdp-based dialogue systems. In *9th Annual Conference of the International Speech Communication Association, 2008*. Interspeech, pages 1169–1172.

Alfred Kobsa. 1990. User modeling in dialog systems: Potentials and hazards. *AI & society* 4(3):214–231.

M. LaMar, R. S.J.D Baker, and Samuel Greiff. 2017. Methods for assessing inquiry: Machine-learned and theoretical. In *Design recommendations for Intelligent Tutoring Systems: Assessment*, volume 5.

Michelle LaMar. 2018. Markov decision process measurement model. *Psychometrika* .

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing* 8(1):11–23.

Ian McGraw, Chia-ying Lee, I Lee Hetherington, Stephanie Seneff, and Jim Glass. 2010. Collecting voices from the cloud. In *LREC*.

Joel Michael, Allen Rovick, Michael Glass, Yujian Zhou, and Martha Evens. 2003. Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments* 11(3):233–262.

Robert J. Mislevy, Russell Almond, Lou Dibello, Frank Jenkins, Linda Steinberg, Duanli Yan, and Deniz Senturk. 2002. Modeling conditional probabilities in complex educational assessments. CSE tech. rep., The National Center for Research on Evaluation, Standards, Student Testing, Cen- ter for Studies in Education, University of California, Los Angeles, Los Angeles, CA.

Natalie K Person, AC Graesser, L Bautista, EC Mathews, Tutoring Research Group, et al. 2001. Evaluating student learning gains in two versions of autotutor. *Artificial intelligence in education: AI-ED in the wired and wireless future* pages 286–293.

Olivier Pietquin. 2005. *A framework for unsupervised learning of dialogue strategies*. Presses univ. de Louvain.

Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education* 16(2):171–194.

Anna N. Rafferty, Michelle M. LaMar, and Thomas L. Griffiths. 2015. Inferring Learners' Knowledge From Their Actions. *Cognitive Science* 39(3):584–618. https://doi.org/10.1111/cogs.12157.

Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Alexei V Ivanov, Keelan Evanini, Zhou Yu, Eugene Tsuprun, and Yao Qian. 2016. Bootstrapping development of a cloud-based spoken dialog system in the educational domain from scratch using crowdsourced data. *ETS Research Report Series* pages 1–7.

Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Alexei V Ivanov, Zhou Yu, Yao Qian, and Keelan Evanini. 2017. Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System. In *Multimodal Interaction with W3C Standards*, Springer, pages 295–310.

Emmanuel Rayner, Ian Frank, Cathy Chua, Nikolaos Tsourakis, and Pierrette Bouillon. 2011. For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language call application .

Vasile Rus, Sidney DMello, Xiangen Hu, and Arthur Graesser. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine* 34(3):42–54.

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing* 20:53–76. https://doi.org/10.1016/j.asw.2013.04.001.

Alexandria Katarina Vail and Kristy Elizabeth Boyer. 2014. Adapting to personality over time: examining the effectiveness of dialogue policy progressions in task-oriented interaction. In *Proceedings of the 15th Annual SIGDIAL Meeting on Discourse and Dialogue*. pages 41–50.

Kurt VanLehn, Collin Lynch, Linwood Taylor, Anders Weinstein, Robert Shelby, Kay Schulze, Don Treacy, and Mary Wintersgill. 2002. Minimally invasive tutoring of complex physics problem solving. In *Intelligent Tutoring Systems: 6th International Conference, ITS 2002, Biarritz, France and San Sebastian, Spain, June 2-7, 2002. Proceedings*. Springer, pages 43–55.

Lawrence G. Weiss and Patricia Zureich. 2008. A primer on psychometrics: The important points for speech–language pathologists. *Pearson Education, Inc.* 24:1–12.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language* 21(2):393–422.

Steve Young. 2006. Using POMDPs for dialog management. In *Spoken Language Technology Workshop, 2006. IEEE*. IEEE, pages 8–13.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language* 24(2):150–174.