PSYCHOMETRIKA—VOL. 83, NO. 1, 67–88
MARCH 2018
https://doi.org/10.1007/s11336-017-9570-0

MARKOV DECISION PROCESS MEASUREMENT MODEL

MICHELLE M. LAMAR
EDUCATIONAL TESTING SERVICE

Within-task actions can provide additional information on student competencies but are challenging to model. This paper explores the potential of using a cognitive model for decision making, the Markov decision process, to provide a mapping between within-task actions and latent traits of interest. Psychometric properties of the model are explored, and simulation studies report on parameter recovery within the context of a simple strategy game. The model is then applied to empirical data from an educational game. Estimates from the model are found to correlate more strongly with posttest results than a partial-credit IRT model based on outcome data alone.

Key words: latent-trait model, cognitive model, performance assessment.

Complex performance tasks often provide potentially useful information in the series of problem-solving actions taken to complete the task; however, these data are challenging to model with traditional psychometric models. Continuous latent-trait measurement models rely on a meaningful mapping between performance indicators and the latent constructs which are the target of inference. For most item response theory (IRT) models, this mapping is provided by expert judgment via a scoring matrix or answer key, where all possible responses to each discrete item are assigned a score for one or more of the latent-trait dimensions. When performances are more complex, however, meaningful student actions are embedded in a larger context which make their ultimate value dependent upon both the current state of the problem and the student's subsequent actions as they implement a strategic plan. For example, if one wishes to measure ability in a board game, the final outcome is easy to score, but the value of a particular move might vary greatly depending upon the current configuration of the board and potential future moves. To utilize the information contained in such within-task actions, a model is needed that can both provide action scores and account for the dependence inherent in the action sequencing.

For many complex tasks, a dynamic problem state can be defined that includes all of the factors that are relevant to the within-task decisions which make up the complex performance. In board games, for example, the configuration of the pieces on the board is often a complete description the problem state. The problem state is then sufficient to give the within-task actions both value and independence, and as such, the state can be considered analogous to an item in a traditional assessment framework. Thus, rather than identifying responses by sequence, they can be associated with the problem state in which they were chosen. The number of possible "items" would then be as large as the state space, and each student would only have responses for a small number of them. Given this approach, the probability of student j taking an action in a particular state s can be modeled using an IRT framework by including the scoring parameter within the model as in the nominal response model (Bock, 1972),

$$p(x_{sj} = r|\theta_j) = \frac{\exp(b_{rs}\theta_j + \xi_{sr})}{\sum_{m=1}^{R_s}\exp(b_{ms}\theta_j + \xi_{sm})}, \quad \sum_{r=1}^{R_s}\xi_{sr} = 0, \sum_{r=1}^{R_s} b_{sr} = 0, \theta_j \sim N(\mu, \sigma^2), \quad (1)$$

Electronic supplementary material The online version of this article (doi:10.1007/s11336-017-9570-0) contains supplementary material, which is available to authorized users.

Correspondence should be made to Michelle M. LaMar, Educational Testing Service, Princeton, NJ, USA. Email: mlamar@ets.org

where b_{rs} is the score parameter for response r in state s, R_s is the total number of response alternatives possible in state s, ξ_{sr} is the intercept for response r in state s, and θ_j is the ability of person j. Estimating this model requires estimating $2R_s - 2$ parameters for each state contained in the data. If the score parameters are fixed by expert judgment, as is traditionally done in a partial-credit model, the number of parameters would be reduced to $R_s - 1$ per state, but scores would need to be specified for every action in every state. To put this in perspective, a simple strategy game such as tic-tac-toe has 5478 legal game states (board positions), with each state allowing 2–9 possible actions. Specifying scoring for each state–action pair by hand would be a monumental task. Estimating the remaining parameters would require massive amounts of data, given that each game would only provide data for at most nine game states.

Instead, this paper proposes explicitly modeling the student's decision making process to express the scoring parameters in terms of a smaller, more tractable, set of parameters, similar to the approach taken in an LLTM (Fischer, 1973). For this purpose, a Markov decision process is used, which links the probability of choosing a particular action in a particular state to the likelihood of achieving a predefined goal, such as succeeding in the task at hand. This paper develops a measurement model for complex assessment tasks by combining the IRT approach with a cognitive model based on the Markov decision process. The remainder of the paper is organized as follows: Sect. 1 reviews the Markov decision process, while Sects. 2 and 3 describe its formulation and estimation as a measurement model. Section 4 explores psychometric properties of the model. Section 5 presents simulation studies that demonstrate parameter recovery, and Sect. 6 briefly describes an application to real-world data, with conclusions and discussion in Sect. 7.

1. Markov Decision Process as a Cognitive Model

A Markov decision process (MDP) is a model for decision making in the presence of uncertainty based on a longitudinal cost–benefit analysis (Puterman, 1994). MDPs have been used extensively in artificial intelligence and robotics to choose optimal actions in stochastic, dynamic situations (Mnih et al. 2015; Russell & Norvig, 2009) and in economics to model individual choice strategies (Rust, 1994).

Formally, an MDP is defined by $\{S, A, T, R, \gamma\}$ where S is the set of possible states of the system and A is the set of possible actions. T represents the transition model, $p(s'|s, a)$, the probability of transitioning to a state s' given that action a was taken in state s. R corresponds to the reward structure $r(s, a, s')$ which specifies the immediate reward for taking action a in state s and entering state s', while $\gamma \in [0, 1]$ is the discount parameter, representing 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:

$$Q(s, a) = \sum_{s' \in S} p(s'|s, a) \left(r(s, a, s') + \gamma \sum_{a' \in A} p(a'|s') Q(s', a') \right), \tag{2}$$

where $p(a|s)$ is the decision rule, or policy, by which actions are chosen given a particular state. The Q function essentially assigns a value to each action in each state. In Eq. (2), $\sum_{a' \in A} p(a'|s') Q(s', a')$ is the expected value of the next state, s', 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. Note that the function is recursive, as the value of a state is defined using the Q function itself. The Q function can be calculated using dynamic programming (Howard, 1960).

In robotics, the MDP is used to solve for an optimal policy $\pi(s)$, which is the set of actions for a given state s that maximize the expected total rewards (Puterman, 1994). More recently, Markov decision processes have been used as a cognitive model to describe not only human decision making, but also people's ability to infer the goals and beliefs of others (Baker, Saxe & Tenenbaum, 2009). Baker, Saxe, and Tenenbaum (2011), describe a "Bayesian theory of mind" in which cognition is modeled as a partially observed MDP. They hypothesize that people act based on their beliefs, modeled by the state space, action set and transition functions, and in accordance with their desires, which are modeled by the reward structure. When modeling human decision making, the policy is not assumed to be optimal, as humans make mistakes. Frequently, a Boltzmann policy is used (Baker et al. 2009):

$$p(a|s) \propto e^{\beta Q(s,a)}, \tag{3}$$

where $\beta \in [0, \infty)$ represents the decision maker's capability to choose actions that will result in higher total rewards. As β increases, the probability of taking the action with the highest Q value, i.e., $\pi(s)$, increases. When β goes to zero, the action probabilities become equal, and actions are selected uniformly at random from the action set.

Note that under this model the decision maker is at worst performing randomly. As a cognitive model, the MDP specifies the individual's actual goals and beliefs, and it is assumed that the individual's actions are consistent with those goals and beliefs based on the principle of rationality (Baker et al. 2009). Thus, while an individual might make mistakes in the pursuit of their goals, represented by lower β values, they will not consistently act contrary to their interests, based on their understanding of the situation. The subjective quality of the model allows it to be used for making inferences about different elements of an agent's cognition, based upon their actions (Baker et al. 2011; Rafferty, LaMar & Griffiths, 2015). In particular, inverse reinforcement learning utilizes MDPs to infer discrete goals based on action traces by estimating the most probable values of the reward function (Ng & Russell, 2000), while inverse planning algorithms infer student understanding of the effects of their actions by estimating parameters in the transition function (Rafferty et al. 2015).

2. Markov Decision Processes for Assessment

The MDP has been primarily used as predictive model, to guide decision making in AIs or to examine the extent to which human decision making matches that predicted by the MDP. Given the success of the MDP as a cognitive model (Baker et al. 2011), a natural extension is to use the model for inference about individual humans. This paper introduces the Markov decision process measurement model (MDP-MM), a latent-trait model through which person traits of interest are inferred based on performance records consisting of actions and problem states in which the actions were taken. As a measurement model, the validity of the inferences made depends on the extent to which the underlying model assumptions are justifiable. The model parameter space and assumptions will be discussed next, followed by a more formal definition of the measurement model.

2.1. Parameters, Constraints and Assumptions

The parameter space of the MDP-MM can be divided into four separable components of the cognitive model (Table 1). The *goal* represents what is to be achieved and/or avoided and thus is categorical, while *motivation* is one or more continuous parameters quantifying the amount of effort put into pursuing the goal. *Understanding* represents the model of the task, which includes

TABLE 1.
Parameter space for the MDP-MM.

Name	Symbol	Type	Interpretation
Goal	R_{g}	Discrete	Final state attempting to achieve
Motivation	R_{m}	Continuous	Degree of effort applied to the task
Understanding	T_h	Discrete	Understanding of task dynamics
Capability	β	Continuous	Decision making capability

the state space, the action set, and the transitions functions. While elements of *understanding* might be arguably continuous (e.g., the probability of a particular outcome in the transition function), in this parameterization *understanding* is a set of discrete hypotheses which might represent misconceptions or differing knowledge about the task. Finally, *capability* is a continuous parameter that represents ability to optimize decision making given the rest of the model.

The assumptions made by the model are largely dependent upon how these parameters are constrained. When all parameters are allowed to freely vary by person and by time, the model becomes completely subjective as it now represents the current internal cognitive model of the decision maker. Goals are defined to be the person's goals; rewards are, by definition, the internal-valued costs and benefits; and the state space, action set and transitions all represent the person's mental model of the world. This unrestricted model is the idealized cognitive model, and it rests on the assumption of rationality, which is that the decision maker will attempt to take actions leading to higher total subjectively valued rewards (Baker et al. 2009). Note that the model does not assume the decision maker can internally calculate the Q function, but only that they weigh the benefits of future outcomes by an internally generated probability of that outcome coming to pass. While such a completely unrestricted model is unlikely to be identifiable, as part of a larger MDP measurement framework, this model can be seen as the basis for specific measurement models that target particular parameters for person-level inference.

At the other extreme, the most constrained measurement model fixes all parameters other than β, which can be estimated at the person level. These constraints imply numerous additional assumptions about the decision makers. First, as the model is time invariant, it assumes that learning does not take place during the task. While violations of this assumption are probable, it is a common assumption in psychometric models and is unlikely to be problematic if the tasks are relatively short and not instructional in nature. Second, the fully constrained model assumes that decision makers act based on the set goals and motivation. Again, this is not dissimilar to the assumptions of IRT models, and violations can be minimized with clear task instructions and sufficient performance consequences. Third, the constrained model assumes that all decision makers share a single, correct understanding of the problem space, action set and state transitions. This assumption is reasonable in highly structured tasks such as a solitaire board game, but is likely to be violated in tasks with hidden dynamics or subtly probabilistic outcomes. When the modeled goals and understanding are fixed by expert opinion, however, this can be seen as parallel to providing a scoring matrix, as the fixed goals and understanding represent the ideal behavior against which we are measuring the test takers. In this model, β measures not only strategic decision making, but also the extent to which behavior conforms to the set R_{g} and T_h. Motivation can also be set at a fixed high value, or it can be estimated at the population level to allow for the fact that there is no "correct" level of motivation, but there may be fairly homogeneous motivation in a group of students who experience the assessment under the same conditions. This highly constrained measurement model will be further developed in this paper, while expansions of the model that loosen these constraints will be discussed at the discussion section.

A few other assumptions and limitations of the MDP-MM should be noted. First, as a Markov model, it is assumed that all factors which influence the value of the action choices can be fully specified within the state definitions. Further, as a practical consideration, the size of the state space cannot grow beyond the limits of the available computing platforms. Thus, there is a limit to the complexity that can currently be modeled, with 500,000 states approaching the maximum that can currently be modeled.

2.2. *Markov Decision Processes for Capability Assessment*

This paper will focus primarily on use of the MDP-MM to estimate the parameter β, as a measure of a student's capability to optimally solve a specific problem. In previous work using the MDP model for inference (Baker et al. 2011; Rafferty et al. 2015), person-specific goals and beliefs were estimated, but β was assumed to be common across participants, and was either fixed or estimated at the population level. Here it is assumed to be person specific and so is notated as β_j. The formulation of the Q function remains as in Eq. (2), except that the dependency upon the capability parameter β_j is explicitly noted. The conditional probability of student j selecting action a when in state s now becomes

$$p(a|s, \beta_j) = \frac{\exp(\beta_j Q(s, a|\beta_j))}{\sum_{a' \in A} \exp(\beta_j Q(s, a'|\beta_j))}. \tag{4}$$

Structurally, Eq. (4) is very similar to the nominal response model (Eq. 1) and is arguably part of the family of IRT models known as divide-by-total models (Thissen & Steinberg, 1986). The $Q(s, a)$ value is analogous to a score parameter, b_{rs} in that it assigns value to the action choice, while β_j is analogous to θ_j as a person parameter that affects the chance of taking higher-scoring actions. β_j is the ability to choose the more highly valued options across different states.

Different from an IRT framework, however, even with fixed T and R values, the action values can differ between persons, as the Q values themselves depend on β_j. When β_j is lower, the MDP model gives more weight to the possibility that a mistake will be made in future actions. For example, an amateur mountain climber might correctly choose a longer but safer path because they realize that they are not skilled enough to safely scale a particular cliff. A more experienced climber could choose the shorter, more dangerous path, again correctly, because they know they have the skills required to make the climb. In this example, we can assume that the reward structure and the transition functions are understood by both, but part of the decision making of each climber involves evaluating the probability that they will make mistakes in future states. For the MDP-MM, $Q(s, a|\beta_j)$ is referred to as the *action value*, which is dependent upon both person and state.

Another difference from standard IRT formulations is that Eq. (4) lacks an intercept parameter, which in the IRT framework represents item difficulty for a dichotomous model or response attractiveness in a partial-credit or nominal model. The lack of intercept implies that for all states, as β_j goes to zero, the probability of selecting any action a goes to $1/R$ where $R = |A_s|$. While the lack of intercept suggests that the model ignores differences in decision difficulty, within the MDP framework decision difficulty is instead represented by the contrast among the Q values for the available actions of a given state. When a decision is "easy," one action will have a much higher Q value than the others, making the selection of the correct action quite probable. For a more challenging decision, the distinction between the actions will be more subtle and the Q values will be closer together. The implications of interpreting Q value differentials for decision difficulty is further explored in Sect. 4.

While the MDP model cannot allow β_j to be negative, as the dynamic calculation of the Q function would no longer converge, performances which display less-than-random-chance behavior would not only result in near-zero estimates of β_j, but they would also display poor model fit, providing an indication that the decision maker held different goals or understanding from the fixed parameters. An example of this dynamic will be found in the application study (Sect. 6).

3. Estimation

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

$$O_j = \{(s_{1j}, a_{1j}), (s_{2j}, a_{2j}), \ldots (s_{T_j j}, a_{T_j j})\}, \tag{5}$$

where T_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

$$p(O_j|\beta_j) = \prod_{t=1}^{T_j} p(a_{tj}|s_{tj}, \beta_j) = \prod_{t=1}^{T_j} \frac{\exp\left(Q(s_{tj}, a_{tj}|\beta_j)\beta_j\right)}{\sum_{a' \in A} \exp\left(Q(s_{tj}, a'|\beta_j)\beta_j\right)}. \tag{6}$$

The model is estimated by taking the β_j parameter as having a log-normal distribution, $\ln N(\mu, \sigma^2)$ across persons. A variable is log-normally distributed when a log transformation of the variable would be normally distributed; thus, β_j can be alternatively defined as $\beta_j = \exp(\lambda_j)$ with $\lambda_j \sim N(\mu, \sigma^2)$. The use of the log-normal distribution ensures that β_j is restricted to be nonnegative, as is required for solving the MDP Q function, but it also makes sense conceptually. The log-normal distribution is found naturally as the distribution of growth metrics (Limpert, Stahel & Abbt, 2001), in particular when growth is best modeled by a multiplicative rather than an additive process. Log-normal ability distributions have also been used previously in psychometric models to achieve a lower bound to the probability of response selection without the introduction of guessing parameters (Bradshaw & Templin, 2014).

The set of all model parameters is defined to include the capability distribution parameters, μ and σ, along with the Q function parameters, R_{m}, which constitute the parameters needed to define motivation within the reward structure. The model parameters are estimated jointly using marginal maximum likelihood (MML), marginalizing over the β_j distribution. The marginal likelihood for the model parameters is

$$L(\mu, \sigma, R) = \int_{\beta_j} \prod_{j}^{N} p(O_j|\beta_j; R) G(\beta_j; \mu, \sigma) d\beta_j, \tag{7}$$

where $G(\beta_j; \mu, \sigma)$ is the log-normal probability distribution. This likelihood cannot be evaluated analytically. Not only is the integral intractable, but the Q-function on which it depends must be calculated through iterative approximation. To evaluate the integral, Gaussian quadrature is used for integration over $\lambda_j = \ln(\beta_j)$ with nine quadrature points. The MDP Q-function is calculated for each candidate set of model parameters using policy iteration methods (Howard, 1960) in which the state values and action probabilities are iteratively updated until they are changing less than a specified convergence criterion.

After the model parameters have been estimated, β_j parameters are predicted using empirical Bayesian estimation, with the prior set to the empirically estimated distribution $\ln N(\hat{\mu}, \hat{\sigma}^2)$. The point estimates are taken as the maximum a posteriori (MAP) estimates, which correspond to the β_j values at which the posterior distribution is maximized,

$$\hat{\beta}_j = \underset{\beta_j}{\operatorname{argmax}}\, p(O_j|\beta_j) G(\beta_j; \mu, \sigma). \tag{8}$$

All of the estimation code was custom-written using the C++ programming language.

4. Psychometric Properties of the Model

4.1. Monotonicity and Decision Difficulty Ordering

As a measure of capability within a problem space, we desire specific relationships between β_j and the probability of selecting an "optimal action" for any given state. Recall that $\pi(s)$ is defined as the set of optimal actions for state s. If a person selects action a while in state s, the probability that they selected an optimal action can be defined as $p(a \in \pi(s)|s, \beta_j)$.

For a continuous measurement model, one expects monotonicity between β_j and $p(a \in \pi(s)|s, \beta_j)$, meaning that as β_j values increase, the probability of selecting an optimal action should be non-decreasing. Defining $\Delta Q(s, a'|\beta_j) \equiv Q(s, \pi(s)|\beta_j) - Q(s, a'|\beta_j)$, the log odds of choosing an element of $\pi(s)$ rather than $a' \notin \pi(s)$ can be expressed as

$$\begin{aligned}
\log\left(\frac{p(\pi(s)|s, \beta_j)}{p(a'|s, \beta_j)}\right) &= \log\left(\frac{\exp\{\beta_j[Q(s, a'|\beta_j) + \Delta Q(s, a'|\beta_j)]\}}{\exp\{\beta_j Q(s, a'|\beta_j)\}}\right) \\
&= \log\left(\frac{\exp\{\beta_j Q(s, a'|\beta_j)\}\exp\{\beta_j \Delta Q(s, a'|\beta_j)\}}{\exp\{\beta_j Q(s, a'|\beta_j)\}}\right) \\
&= \beta_j \Delta Q(s, a'|\beta_j).
\end{aligned} \tag{9}$$

For a state s whose value does not depend upon future actions, i.e., a terminal state, the Q function does not depend upon β_j and the log odds of choosing $\pi(s)$ increases as β_j increases. For any other state, ΔQ depends upon β_j because β_j affects the probability of selecting optimal actions in the future. It can be shown that increasing β_j always increases the value of future states, but the interplay of the transition probabilities and the reward structure complicates the proof that ΔQ always increases as β_j increases in every state. Instead, we suggest that the weaker criterion of convergence to an optimal action as β_j increases should suffice to ensure that β_j can be interpreted as a measure of the capability to find an optimal solution. Formally then we require that as $\lim_{\beta_j \to \infty} p(a \in \pi(s)|s, \beta_j) = 1.0$.

This requirement is easily satisfied by the MDP-MM. Note that the probability of selecting an action from a finite set of actions A is

$$p(a|s, \beta_j) = \frac{\exp[\beta_j Q(s, a|\beta_j)]}{\sum_{a' \in A} \exp[\beta_j Q(s, a'|\beta_j)]}.$$

The set of actions $\pi(s)$ is defined as actions which maximize $Q(s, a|\beta_j)$. An element of this set is referred to as $a_{\pi(s)}$. For each $a' \notin \pi(s)$, there exists a constant $c_{a'} > 0$ such that

$$Q(s, a'|\beta_j) = Q(s, a_{\pi(s)}|\beta_j) - c_{a'}.$$

Out of the action set A, the number of optimal actions in state s is defined as $N_\pi = |\pi(S)|$. Now the probability of selecting a particular $a_{\pi(s)} \in \pi(s)$ is

$$
\begin{aligned}
p(a_{\pi(s)}|s,\beta_j) &= \frac{\exp\left[\beta_j Q(s,a_{\pi(s)}|\beta_j)\right]}{\sum_{a'\in A}\exp\left[\beta_j Q(s,a'|\beta_j)\right]} = \left\{\sum_{a'\in A}\frac{\exp\left[\beta_j Q(s,a'|\beta_j)\right]}{\exp\left[\beta_j Q(s,a_{\pi(s)}|\beta_j)\right]}\right\}^{-1} \\
&= \left\{N_\pi + \sum_{a'\notin\pi(s)}\frac{\exp\left[\beta_j Q(s,a_{\pi(s)}|\beta_j) - \beta_j c_{a'}\right]}{\exp\left[\beta_j Q(s,a_{\pi(s)}|\beta_j)\right]}\right\}^{-1} \\
&= \left\{N_\pi + \sum_{a'\notin\pi(s)}\frac{\exp\left[\beta_j Q(s,a_{\pi(s)}|\beta_j)\right]}{\exp\left[\beta_j Q(s,a_{\pi(s)}|\beta_j)\right]\exp\beta_j c_{a'}}\right\}^{-1} \\
&= \left\{N_\pi + \sum_{a'\notin\pi(s)}\frac{1}{\exp\beta_j c_{a'}}\right\}^{-1}.
\end{aligned}
$$

As $\beta_j \to \infty$, $\frac{1}{\exp(\beta_j c_{a'})} \to 0$ for each $a' \notin \pi(s)$. Thus,

$$
\lim_{\beta_j\to\infty} p(a_{\pi(s)}|s,\beta_j) = \frac{1}{N_\pi}.
$$

The probability of selecting an action that is contained in the set of optimal action is

$$
p(a \in \pi(s)|s,\beta_j) = \sum_{a_{\pi(s)}} p(a_{\pi(s)}|s,\beta_j),
$$

so

$$
\lim_{\beta_j\to\infty} p(a \in \pi(s)|s,\beta_j) = \sum_{a_{\pi(s)}}\frac{1}{N_\pi} = 1.
$$

While it is reassuring that increased capability converges to optimal action choice, care must be taken to ensure that the mathematical definition of optimal matches the intended best path(s) through the problem. Using the generative model to simulate action choices can be quite useful to determine if the transition structure allows unintended shortcuts to higher rewards.

Another interesting relationship in psychometric models is between item difficulty and the probability of optimal response. As the Q values determine decision difficulty in the MDP-MM, their effect on predicted performance as a function of β_j will be examined. First, it should be noted that the decision difficulty is not affected by the overall magnitude of the Q function. This is clear from the fact that one could add an arbitrary constant C to all Q values in Eq. (4), and it would factor out of both the numerator and denominator, thus canceling. Decision difficulty, then, is determined by the difference between the Q values of the available options, as defined by $\Delta Q(s,a'|\beta_j)$ above. Given a decision with only two options and for which $\Delta Q(s,a'|\beta_j) = \Delta Q(s,a')$, the decision characteristic curves for four such decisions of differing values of $\Delta Q(s,a')$ are compared in Fig. 1. High-contrast decisions, such as $\Delta Q(s,a') = 4.0$, provide clear distinction between capabilities near zero, but do little to distinguish the higher-capability students. The low-contrast decisions, on the other hand, result in smaller differences in the action probabilities, but continue to distinguish students into the high β_j values.

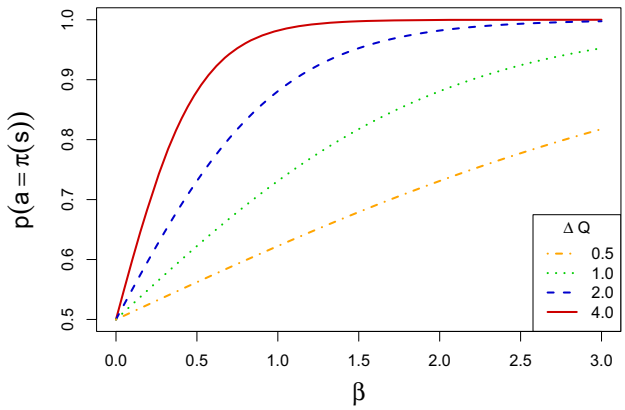

FIGURE 1.
The probability of selecting an optimal response as a function of capability, β_j, for 2-option decisions of differing $\Delta Q(s, a')$.

4.2. *Parameter Interpretation*

The probability of selecting a particular action (Eq. 4) is affected both by the value of β_j and by the differences between the Q values for the available actions. The Q value differences are themselves affected by two parts of the model, the β_j values, as they affect the probability of future optimal action, and the reward values which provide a scale for Q. Note that the capability parameter β_j has two mechanisms by which it affects the action probabilities. We refer to the effect of β_j as it acts through the Q function as the *indirect* effect of β_j, contrasted with the *direct* effect of β_j as a parameter in the decision model. Each of these three components can be interpreted cognitively. The direct effect of β_j models the student's ability to make a correct decision given their understanding of the problem. The indirect effect of β_j models the student's ability to understand how their actions will affect future states. Note that the farther away the goal is, the more diluted the Q values become for a low β_j student. This maps to an intuitive understanding that higher-capability students would be able to "read" farther into the problem than lower-capability students. Finally, reward structure magnitudes, R_m, model the motivation of the students. As reward differences increase, the Q values differences increase, leading to an increased probability of choosing an optimal action. Further, a larger move penalty will cause lower-capability students to gravitate to shorter paths through the problem space, even if they do not result in the highest final reward.

4.3. *Identification*

The MDP-MM has two sources of potential identification problems. First, as in other divide-by-total models, the logit is invariant to translation, meaning that an arbitrary constant could be added to all values $\beta_j Q(s, a|\beta_j)$ for a given decision without changing the decision probabilities. Second, because of the multiplicative relationship between β_j and $Q(s, a|\beta_j)$, an arbitrary factor c could also be introduced to both values as

$$\beta_j Q(s, a|\beta_j) = (c\beta_j)\left(\frac{Q(s, a|\beta_j)}{c}\right). \tag{10}$$

Given the log-normal distribution, multiplying β_j by a constant is equivalent to an additive translation of the log-transformed normal variable λ_j,

$$c\beta_j = c\exp(\lambda_j) = \exp(\ln c + \lambda_j) = \exp(\lambda_j + c'), \tag{11}$$

thus resulting in a shift of the μ parameter but not affecting σ. Note that multiplying β_j by a factor c would also affect $Q(s, a|\beta_j)$ in many states due to the dependence on β_j. Thus, this is not a pure identification problem, but may result in weak identifiability, causing problems with the estimation algorithms.

To produce a clearly identified model, therefore, either the scale and location of the β_j distribution or the scale and location of the $Q(s, a|\beta_j)$ distribution will need to be constrained. As the β_j distribution is more readily interpretable as characterizing a population, constraints are applied instead to the $Q(s, a|\beta_j)$ values. First, the maximum $Q(s, a|\beta_j)$ is fixed to give the location constraint. This is straightforward in most MDP specifications, as generally there is an end goal with a large payoff which drives the process. Next, the scale of $Q(s, a|\beta_j)$ is constrained by fixing an interval within the range of possible values. Again, depending upon the parameterization of R, this can usually be easily handled by setting a second parameter within the reward structure.

It should be noted that the identifiability of a model is not a mere statistical technicality. Model identification problems indicate the limits of the inference that can be made from data. The linear identification problem, present in many IRT-based models, indicates that we cannot infer absolute student capability, but only relative student capability. The second identification problem is interesting because it involves an interplay between the reward parameters and the capability parameter. If these were each single parameters, one would conclude that capability and motivation cannot distinguished. In fact, most educational assessment models are unable to distinguish between capability and motivation. While motivation is not even parameterized in common assessment models, there is an implicit assumption that either students are highly motivated so that ability is being measured, or that the construct being measured is actually the combination of ability and motivation. In particular, the highly motivated low-ability student and the poorly motivated high-ability student are generally indistinguishable. An MDP model, however, might be able to make that distinction if there exist more than two reward parameters. The ability of the MDP-MM to distinguish between capability and motivation will be evaluated in the second simulation study in the next section.

5. Simulation Studies

Two simulation studies evaluate the performance of the MDP-MM by examining parameter recovery using the peg solitaire game, which is explained below. Study 1 evaluates basic recovery of the population capability parameters, μ and σ, and the student capability parameter, β_j, under ideal conditions. For these simulations, the model reward parameters, R, are all fixed at favorable, consistent values for both the generating and estimating models. Study 2 evaluates the ability to distinguish capability and motivation at the population level by estimating both the capability distribution and R_{m}, the reward parameter relating to motivation, in four distinct simulated samples. For both studies, the model parameter and capability estimates were compared to the generated values using the bias and root-mean-squared error (RMSE). For recovery of β_j, error metrics for $\log(\hat{\beta}_j)$ are used to make the metrics comparable to more common normally distributed latent-trait parameters.

5.1. Peg Solitaire MDP

The peg solitaire game consists of a board with holes, some of which are filled with pegs. Legal moves involve one peg jumping over an adjacent peg into an empty hole on the other side, after which the jumped peg is removed from the board. The goal is to leave as few pegs on the

board as possible. The complexity of the game can vary depending upon the size and configuration of the board and starting position.

To describe the peg solitaire game as an MDP, each element of the MDP (state space, action set, transition function, reward function and discount parameter) must be specified for the game. The complete state space includes all possible configurations of the pegs on the board, but for any particular game, the state space can be reduced to all reachable peg configurations given the starting configuration. Similarly, the action set for a particular state is defined as all possible actions from that state which include all legal peg jumps in state s along with the *reset* and *score* actions. The *reset* action resets the board back to the starting state. The *score* action ends the game and assigns the final rewards for the board.

The transition model, $p(s'|a, s)$, is deterministic for this game. A legal move will transition to the board state which has the jumping peg moved over and the jumped peg removed with a probability of 1.0. The *reset* action always transitions to the starting position, while the *score* action does not change the game state. The reward function, described using four parameters (Table 2), assigns high reward to the single-peg end state, decreasing rewards for scored positions that include more pegs while assigning negative reward to peg-move and reset actions. Finally, the discount parameter γ is fixed at 1.0, as the value of winning is not smaller earlier in the game than later in the game.

5.2. *Simulation Design*

Four different game boards were used for the simulated assessment (Fig. 2). These boards varied in game length from the tiny cross board, which can be solved in 5 moves, to the big-L board which requires 13 moves to win (Table 3). Other game complexity metrics including state space size and action set size find the diamond board to have the highest complexity due to a greater number of choices per decision point (Table 3).

Play records are simulated using the MDP as a generating model. For all simulations, R_{win} is fixed at 5.0 and R_{peg} is fixed at 1.0 for identifiability; R_{reset} was fixed to -1.0, and R_{move} was taken to be the motivation parameter, R_{m}. Given study-specific generating parameter values for μ, σ, and R_{move}, student samples are generated as

$$\beta_j = \exp(\lambda_j); \quad \lambda_j \sim N(\mu, \sigma^2), \tag{12}$$

and the Q function is solved for each generated β_j. Starting in the state corresponding to the board's start configuration, game records are generated by drawing an action from the state-specific action probabilities given by Eq. (4). If the *score* action is drawn, the game ends; otherwise, the next state is calculated based on the selected action and the process repeats.

TABLE 2.
Reward parameterization for the peg solitaire MDP.

Parameter	Description	Sim value
R_{win}	Reward for scoring with only one peg left on the board	5.0
R_{peg}	Adjustment to score for each additional peg remaining	-1.0
R_{move}	Reward (cost) for each game move	-0.1
R_{reset}	Reward (cost) for resetting the game	-1.0

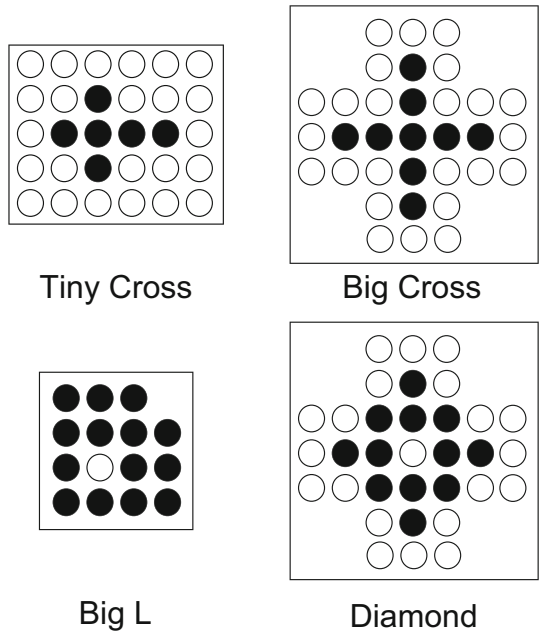

FIGURE 2.
Starting positions for the game boards used in the simulation studies; *black* pegs, *white* holes.

TABLE 3.
Complexity measures for the game boards used in the simulation studies, where the state space is taken to be the states reachable from the start state and the action set contains all possible actions across states.

Board name	Solution path length	Size of state space	Size of action set
Tiny cross	5	22	12
Big cross	8	153	22
Big-L	13	807	30
Diamond	11	5923	70

5.3. Study 1: Recovery of Population and Person Parameters

The first study evaluated the recovery of the population parameters, μ and σ, and the person parameters β_j given an assessment consisting of four tasks, each of the four game boards played once. A single sample of 200 students was simulated from a population with $\mu = 0.0$ and $\sigma = 0.75$. The generating motivation parameters was $R_{\text{move}} = -0.1$. The simulated students each played all four game boards 50 times, giving 1000 assessment records with a total of 4000 games played. This design allows the bias and standard error of the β_j estimates to be more easily evaluated, as well as direct comparison of the information gained from the different game boards. As multiple samples were not simulated, the study does not provide information about how the estimates might be affected by differing sample characteristics.

The population parameters, μ and σ, were estimated using MML, and the individual capability parameters, β_j, were estimated using MAP with the estimated population distribution $\ln N(\hat{\mu}, \hat{\sigma}^2)$ used as the prior. For all estimations, the reward parameters, R, were fixed to the same values as in the generating model. In addition to parameter recovery on the four-game assessment, recovery

TABLE 4.
Mean population parameter estimates (and SE) using fixed reward values for the full four-task assessment and each individual game board.

	μ		σ	
	Bias	RMSE	Bias	RMSE
Full assessment	0.027 (0.013)	0.098 (0.034)	0.057 (0.013)	0.106 (0.061)
Individual games				
Tiny cross	0.017 (0.008)	0.055 (0.023)	0.103 (0.029)	0.217 (0.102)
Big cross	0.028 (0.007)	0.053 (0.024)	0.076 (0.014)	0.121 (0.053)
Big-L	0.019 (0.008)	0.058 (0.034)	0.056 (0.011)	0.093 (0.049)
Diamond	0.012 (0.009)	0.058 (0.033)	0.056 (0.013)	0.105 (0.046)

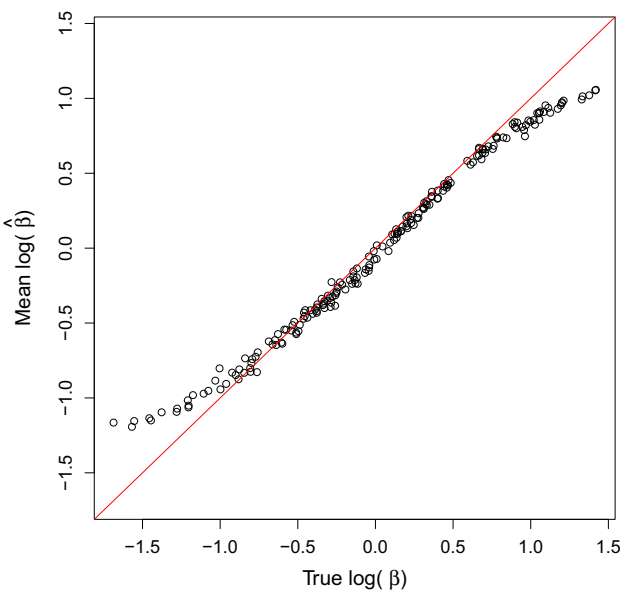

FIGURE 3.
Recovery of person parameters in log space using the full four-task assessment, with the mean taken over 50 replications.

from single-game records is also examined to determine how well capability can be estimated from one game record alone.

5.3.1. Results Population parameter recovery (Table 4) shows a slight positive bias in both $\hat{\mu}$ and $\hat{\sigma}$, but the RMSE for both parameters is small. Note that the log-normal population parameters are equivalent to the transformed normal population parameters and thus, can be directly compared to similar recovery of μ_θ and σ_θ from an IRT study. The four-task assessment does not perform better than individual games in population recovery, and in fact, the RMSE for $\hat{\mu}$ is worse than that estimated from any individual game. In terms of board complexity, the errors for $\hat{\mu}$ remained fairly consistent across different complexities, but σ estimation improves on the more complex boards.

For person parameter recovery, mean $\log(\hat{\beta}_j)$ can be seen to well track true $\log(\beta_j)$ in Fig. 3 (correlation 0.99), with the mean estimate for each simulated student taken across the

TABLE 5.
Recovery of $\log(\beta_j)$ with fixed reward values for the full four-task assessment and for individual game boards (single game play).

Board	Bias	RMSE
Full assessment	−0.034	0.252
Individual boards		
Tiny cross	−0.174	0.542
Big cross	−0.101	0.442
Big-L	−0.096	0.388
Diamond	−0.100	0.403

TABLE 6.
Generating values for the four samples used in simulation study 2.

Sample	Capability	Motivation	μ	σ	R_{move}	R_{reset}	R_{win}	R_{peg}
1 HAHM	High	High	0.5	0.75	−0.05	−1.0	5.0	1.0
2 HALM	High	Low	0.5	0.75	−0.75	−1.0	5.0	1.0
3 LAHM	Low	High	−0.5	0.75	−0.05	−1.0	5.0	1.0
4 LALM	Low	Low	−0.5	0.75	−0.75	−1.0	5.0	1.0

50 replications. The bias of $\log(\hat{\beta}_j)$ is −0.034, and RMSE is 0.25. For the person parameter estimation, the value of multiple tasks is more evident as the $\log(\beta_j)$ errors are significantly reduced for the full four-game assessment compared to estimates based on a single-game record (Table 5). Person parameters are notoriously more difficult to estimate than task-level or population parameters. To put the current results in context, the RMSE can be compared to those attained for $\hat{\theta}_j$ in simulation studies using IRT models. A recent study of parameter recovery using small tests found $\hat{\theta}$ RMSE of 0.44 under their best performing condition which consisted of a 16-item test and 500 simulated students, while the 8-item test achieved an RMSE of 0.56 under best conditions (Svetina et al. 2013). A simulation study using 25 items and the graded response model achieved a $\hat{\theta}$ RMSE of 0.34 under best conditions (Reise & Yu, 1990), and a study comparing a range of test sizes found $\hat{\theta}$ RMSE of 0.56, 0.39 and 0.28 for 15, 30 and 60 items, respectively, using a 2PL IRT model (Hulin, Lissak & Drasgow, 1982). Thus, the present results indicate that the MDP-MM can recover person capability using four tasks with precision similar to a 60-item test modeled with IRT. Furthermore, one notes that given a single game play, the MDP-MM is able to estimate the person capability with precision similar to that of a 15- to 30-item IRT-modeled assessment.

5.4. Study 2: Recovery of Reward and Population Model Parameters

The second simulation study examined separability of capability from motivation at the population level. In particular, the study evaluated distinguishability of highly motivated low-ability populations from poorly motivated high-ability populations.

5.4.1. Design Four samples were generated to simulate populations of differing capability and motivation according to the generating parameters in Table 6. The high-capability populations were generated with $\mu = 0.5$, which gives a median β_j value of 1.65, while the low-capability were generated with $\mu = -0.5$, giving a median β_j of 0.61. Differing motivation was simulated using the move cost, R_{move}, which should be considered relative to the potential gain for taking a

TABLE 7.
Mean number of resets and game score per data set along with mean population parameter estimates by capability and motivation conditions, standard deviation in parentheses.

Sample	# Resets	Game score	μ	σ	R_{move}
1 HAHM	93.2	3.77	0.559 (0.122)	1.016 (0.879)	−0.039 (0.032)
2 HALM	6.5	0.73	0.681 (0.279)	0.743 (0.192)	−0.774 (0.077)
3 LAHM	111.0	0.95	−0.639 (0.482)	0.908 (0.590)	−0.062 (0.047)
4 LALM	28.0	−1.07	−0.400 (0.276)	0.900 (0.514)	−0.774 (0.111)

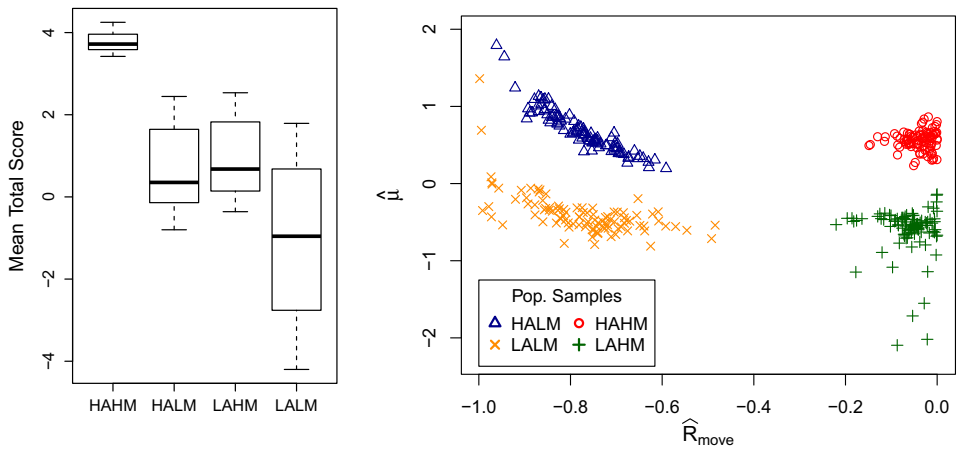

FIGURE 4.
Distribution of per-data set mean game score (*left*) and sample group clustering shown in $\hat{\mu}$ and $\hat{R}_{\text{move}}$ space (*right*). Each condition includes 200 data sets.

single move, R_{peg} which is fixed at 1.0. The high-motivation groups have a very low move cost of −0.05 which would encourage them to keep trying after making a mistake. For the low-motivation students, on the other hand, the cost of a move is half of the expected gain from the move, given perfect future play. For low-capability students, perfect play is unlikely, so this reward value should discourage continued attempts after errors. R_{reset} was fixed to −1.0 in all conditions. For each condition, $N = 200$ students were simulated and each simulated student played each of the four game boards 25 times. The game records were also scored with outcome measures, using the same scoring as implied in the reward parameters, that is 5 points were awarded for scoring the game with a single peg remaining, 4 points if two pegs remained, etc.

5.4.2. Results Performance differences between the four populations are immediately evident in the simulated game records. As predicted, the higher-motivation students persisted longer, resetting the game more often (Table 7). Game score statistics, however, are less distinctive. In particular, the distribution of mean game score from the high-ability low-motivation group (HALM) is not distinguishable from that of the low-ability high-motivation group (LAHM) (Fig. 4, left). Only the high-ability high-motivation (HAHM) group can be distinguished using outcome scores. The estimates from the MDP model, on the other hand, are able to distinguish all groups (Fig. 4, right).

TABLE 8.
Parameter recovery over all replications and all boards by experimental group.

	μ		σ		R_{move}	
Sample	Bias	RMSE	Bias	RMSE	Bias	RMSE
1 HAHM	−0.005	0.122	0.324	0.932	0.011	0.034
2 HALM	0.117	0.301	0.050	0.197	−0.024	0.080
3 LAHM	−0.155	0.504	0.183	0.615	−0.012	0.048
4 LALM	0.084	0.288	0.176	0.540	−0.024	0.113

The mean estimates for the population level parameters (μ, σ and R_{move}), shown in Table 7, are very close to the true values from the sample. While the recovery of R_{move} is fairly consistent across conditions, the recovery of μ and σ vary considerably by experimental condition (Table 8). Even for the best performing condition in this study, the estimation errors are higher than those achieved in study 1. This is to be expected as more parameters are being estimated, but suggests caution in expanding the role of the measurement model to encompass measurement of motivation and perhaps understanding at the person level.

6. Application Study

6.1. Microbes

The MDP-MM was applied to a publicly available educational game, Microbes, in which students learn about cell biology by playing the part of a microbe navigating through increasingly challenging environments (Red Hill Studios, n.d.). On each of the six main game levels, the student receives a description of the environment they are about to enter, is given an opportunity to configure their microbe by "buying" upgrades, and then enters the water tank with their configured microbe, attempting to navigate the environment to reach a marked goal without being eaten or running out of energy. If the student succeeds, they are given ten game tokens and move on to the next level. If they fail, they are encouraged to try again, reconfiguring their microbe if they wish. The recorded data include configuration actions, the decision to enter the tank and the outcome (success or death) of each attempt. A pilot study evaluating the game's utility as an assessment of students' understanding of cellular energy production included a multiple-choice posttest covering similar content.

6.2. Modeling the Microbes Game

The MDP model for Microbes is designed to capture distinctions important to the construct being measured while avoiding excess complexity. Each of the six levels is modeled as a separate MDP task, with information pooled across tasks in the estimation. The state space primarily consists of the number of mitochondria, $m_s \in \{0, ..., 10\}$, and number of chloroplasts, $c_s \in \{0, ..., 10\}$, contained within the microbe, while the environmental variables of availability of food, $F_l \in \{0, 1, 2\}$, and sun, $S_l \in \{0, ..., 3\}$, are fixed by game level, l. During the play of the game, these variables interact to determine the amount of energy available to the microbe and thus heavily influence its chance of survival.

The MDP action set includes *buy-mito* and *buy-chloro* which increment the respective organelle counts in the microbe, and *enter-tank* which occurs when the student decides to enter

the tank with the current microbe configuration. Finally, the *stop* action occurs when the student decides to stop playing the game level. The reward function R is set up with four parameters, the reward for reaching the goal in the tank is R_{win}, while the reward (cost) for a buy action is R_{buy}, for attempting the tank without succeeding is R_{lose}, and for stopping is R_{stop}.

The transition function, $p(s_{t+1}|a_t, s_t)$, for purchase actions is deterministic. If a student buys a mitochondrion, they will transition to the state in which their microbe configuration includes one more mitochondrion, up to the limit of the state space. The result of *enter-tank*, on the other hand, is probabilistic, as the attempt will either succeed or fail based in part on how well the microbe configuration is suited to the environment. The probability of a successful attempt in a given state is calculated based on a logistic function with theoretically fixed coefficients for the utility of mitochondria and chloroplasts in the presence of different environmental factors. Details can be found in LaMar (2014).

6.3. Data

The data come from 238 middle school students who played at least one game level, 148 of whom also completed the accompanying 24 question posttest. The game play data consists of a sequence of action steps, recording the game state preceding the action, the action taken and the result of the action. Actions that exceeded the limit of modeled organelles were removed from the data. Actions that followed a successful attempt on a level before moving onto the next level were also removed as these actions no longer fit the goal-as-winning oriented cognitive model. The remaining data include 4735 actions with the number of actions per student ranging from 1 to 199, median 19. Of the 238 students, 128 played all six levels and 64 played three or fewer levels.

To compare the MDP capability estimates to more traditional measures, the data are also coded based only on the outcome of the *enter-tank* actions, with each level being considered as an "item." As students are allowed to attempt levels multiple times, however, there is no one-to-one correspondence between levels and play results. To deal with these repeated-attempt data, the levels are scored in two different ways. Dichotomous scoring codes only the outcome of a student's first try on each level as either 0 for failure or 1 for success, with all other attempts being ignored. As an alternative, a partial-credit scoring is implemented in which students receive a 3 for winning on their first try, 2 for a second-try win, and 1 for a third-try win. Students who could not win by their third try are given a score of 0 for that level.

6.4. Models

6.4.1. MDP The Microbe MDP model includes reward parameters which must be specified or estimated. Fixing the reward parameters corresponds to a expert-scored model and would result in capability scores that measure students against a knowledgeable, highly motivated expert. This is estimated as model 1. To explore the potential of estimating population-level motivation along with person-level capability, reward parameters are also estimated in model 2. For model identification, two reward parameters are constrained: $R_{\text{win}} = 1$ and $R_{\text{stop}} = 0$. R_{buy} and R_{lose} are estimated at the population level as an indication of the subjective cost that students generally assign to time spent configuring their microbe and to losing their attempt in the tank. To maintain meaning of the parameters, however, the range of R_{buy} is restricted as $R_{\text{buy}} \in (-R_{\text{win}}, 0)$, because values outside of this range would fundamentally change the cognitive model represented by the MDP. Model parameters for all models are estimated using MML as was done for the simulation studies, but for this study, β_j is then predicted using straightforward MLE.

6.4.2. IRT The outcome data (success or failure per level attempt) are modeled using a Rasch model, estimated with the item parameters constrained to sum to zero. The "first-try" outcome

TABLE 9.
Comparison of MDP models. Note that fit metrics for 2' are not comparable to the other models as it was run on a subset of the data.

	Model 1	Model 2	Model 2′
Parameter estimates			
μ	0.051	1.410	1.037
σ	2.503	1.755	1.041
R_{lose}	−0.20[a]	−0.006	−0.035
R_{buy}	−0.10[a]	−0.969	−0.203
Model fit			
−2 log(L)	15461.87	13444.96	11235.18
AIC	15465.87	13452.96	11243.18
np	2	4	4

[a] Fixed parameters.

data are fit using a dichotomous model (IRT FT) and the partial-credit outcome data using a partial-credit model (IRT PC). The posttest data were also fit with a dichotomous Rasch model. All IRT models were run using ACER's ConQuest software (Wu, Adams & Wilson, 1998).

6.5. *Analysis*

The two different MDP models were compared, model 1, which fixed rewards to theoretical values, and model 2 which estimated two reward parameters at the population level. The model fits were compared based on their log-likelihood and AIC fit statistics. The student capability parameters were then estimated using MLE because the MAP estimation was found to be overly sensitive to $\hat{\sigma}$. Pairwise correlations between these capability estimates and the ability estimates from the students' posttest results were examined for validity evidence. The MDP estimates were also compared to those generated by the outcome-only IRT models to determine if the models utilizing the game action data provide any additional information beyond that available in the win/lose data.

6.6. *Results*

Model 2, with estimated reward parameters, produced the best fit by all criteria (Table 9). The estimated distribution of β_j varied widely, with μ estimated near 0 by model 1 and around 1.4 for model 2. The σ parameter was generally estimated to be large.

For model 2, $\hat{R}_{\text{lose}}$ was close to zero, while $\hat{R}_{\text{buy}}$ was nearly the negative of R_{win}. Given such estimates, the MDP would predict that in almost any game state, students would choose to enter the tank rather than buy an organelle, as there would be little cost for attempting and losing compared to the relatively large cost for upgrading the microbe. These estimates suggest that a significant proportion of the sample students are indeed choosing to frequently enter the tank without buying many organelles. The reward estimates, while within the range for what is rational, change the meaning of the MDP as a measurement model as they change which actions are considered optimal, implying that buying organelles reflects poor judgment. Thus for model 2, β_j may no longer serve as a measurement of understanding cell biology. In fact, the correlation of $\hat{\beta}_j$ between models 1 and 2 is −.43. While we would like to use the estimates of R_{buy} to measure the general motivation of the students while also measuring their content understanding with β_j, this methodology cannot be used if the estimated R values fundamentally change the measurement construct.

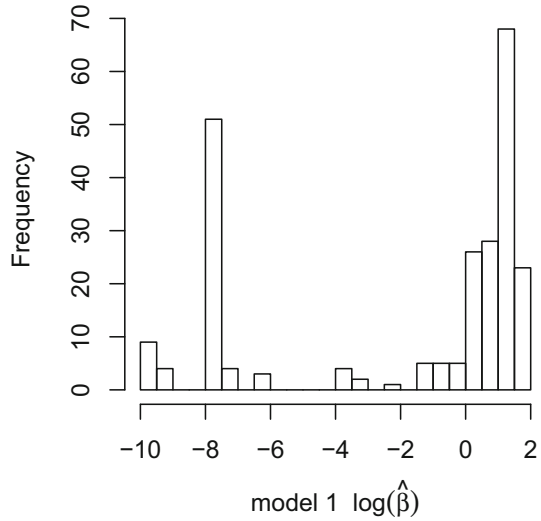

FIGURE 5.
Distribution of $\log(\hat{\beta}_j)$ for the fixed reward models.

For the fixed reward model, model 1, the distribution of the $\log(\hat{\beta}_j)$ values appears bimodal (Fig. 5), though one would expect $\log(\beta_j)$ to be normally distributed. Thus, it seems probable that the sample population comes from two distinct subpopulations. Examination of the play records for the students whose estimates were in the lower clusters revealed that a large proportion of these students never bought any mitochondria or chloroplasts.

Of the original 238 students, 50 never bought a single organelle. These were not students who merely quit early; the median number of levels played in this group was 5 out of the 6 possible, while the median number of tank attempts was 17.5. For model 1, out of the 71 students whose $\log(\hat{\beta}_j)$ was less than -5, 46 were no-buy students. While some of these no-buy students are likely to have been truly confused about the value of the organelles, it seems plausible that many of them were playing based on a cognitive model that does not match our hypothesized model, thus complicating the estimation of the reward parameters.

To test whether the estimation of the R parameters would be feasible in the absence of these possibly off-model students, model 2 was estimated again using only those students who had at least one organelle purchase. Those parameters estimates were then used to predict the individual β_j parameters for the full sample, including the no-buy students. These results are presented on the right in Table 9 and referred to as model $2'$ to distinguish it from the original model estimates. Again, the estimated penalty for losing is much smaller than our theoretical value, while the estimated cost for buying is larger than the fixed value. The buy cost this time is not prohibitive, however, and buying organelles remains a probable option in many game states. The predicted β_j values from this model correlate highly with the fixed rewards model, with the results of model $2'$ correlating with model 1 as $\rho = .99$.

As evidence of validity, the predicted β_j are compared to estimates from the posttest. Table 10 shows that model $2'$ yields the highest correlations with the posttest capability estimates, $\rho = .52$. The IRT models correlate less well than any of the candidate MDP models with the best correlating model having $\rho = .38$, suggesting that the process data do in fact provide relevant information beyond what can be gained from the outcome data on which the IRT models rely. The correlation between model $2'$ and IRT PC was .76, which reflects the strong connection between the actions taken in the game and the final outcome. The model rankings by Spearman correlation, which

TABLE 10.
Correlations between capability estimates from MDP models, IRT models and the posttest IRT model.

	Correlation with post-test	
	Pearson	Spearman
MDP mod 1	.507	.474
MDP mod 2′	.516	.492
IRT FT	.317	.311
IRT PC	.379	.388

does not assume normality, are the same as with the Pearson correlations, with model 2′ doing the best overall and the partial-credit model correlating highest of the two IRT models.

7. Discussion

This paper explores the use of the Markov decision process as the basis for a measurement model in the context of complex strategic problems. As a cognitive model for decision making, the MDP includes elements that correspond to goals, motivation, task-understanding (beliefs) and problem-solving capability. Because these elements each affect action probability in different ways, estimation of motivation and beliefs are naturally separable from an estimation of problem-solving capability. Both goals and beliefs have been estimated using MDP models before (Baker et al. 2011; Ng & Russell, 2000; Rafferty et al. 2015), but in those cases the focus was on classifying agents by their goals and beliefs and the Boltzmann parameter, β, was taken to be a nuisance parameter at most. This present study focused on the Boltzmann parameter as a measure of problem-solving capability and further explored how the MDP might be used to separate the commonly confounded latent traits of capability and motivation.

The simulation studies showed that both population-level and student-level parameters can be reasonably recovered, with a few complex tasks providing as accurate capability estimates as traditional IRT with 60 items. The model was able to accurately estimate both capability and motivation at the population level. In particular, the MDP-MM was able to clearly separate data sets that were generated under "high-ability but low-motivation" conditions from those that were generated under "low-ability and high-motivation" conditions—a separation that was not possible using the task outcome metric alone. The next step in this line of research would be to classify students into subpopulations based on motivation while still estimating their capability at person level.

The application of the MDP-MM to the Microbe game demonstrated both feasibility of model and significant validity evidence in the positive correlations with posttest capability estimates. Further, the fact that the MDP capability estimates correlated more highly than the best IRT model's estimates suggests that the MDP model, and the process data on which it relies, can yield more information about student competency than outcome data alone. It should be noted, however, that with the highest correlation at .52, all of these model estimates correlate less strongly than one would expect from an alternate form of assessment. As the game in question was not originally created as an assessment, it is likely that the quality of information contained in the performances could be improved through a more principled assessment design process (Mislevy, Behrens, Dicerbo, Frezzo & West, 2012).

The estimation of model parameters proved to be fairly sensitive to correct specification of the cognitive model over the full population. While this would be problematic if the model was to

be indiscriminately applied to large populations, the sensitivity can also be seen as a strength as it provides both a person-level validity check and an opportunity to use the model diagnostically. The straightforward application of the constrained MDP-MM is most appropriate with well-defined tasks in which the goals and task mechanics can be reasonably assumed to be known and shared by all participants. A logical next step, however, is to extend the model to allow for multiple cognitive approaches through the estimation of T_{hj} or R_{gj} using a full mixture model. This extended MDP-MM could be fruitful in more diagnostic applications where particular misconceptions might be identified along with an estimation of generalized problem-solving ability (Rafferty et al. 2015). As it is unlikely that all student cognitive models can be anticipated in a real-life task, attention to person-fit metrics will also be needed to distinguish students for whom none of the candidate models are a good fit.

This study focused on the use of the MDP-MM in modeling game play; however, the model is well suited for other multi-step complex problem-solving tasks. Any task with a discrete action set and a discrete state space has the potential to be modeled as an MDP. With the current emphasis on performance tasks brought by the Common Core curriculum and the Next Generation Science Standards, complex tasks are increasingly included in major assessments (National Research Council, 2014). Models that can utilize the within-task performance information they provide are needed to produce more valid and reliable measurement inferences. The MDP-MM has the potential to provide such measures.

Acknowledgments

The author would like to thank Robert Hone and Matt Silberglitt for sharing their data from the Microbes study, the collection of which was supported by the National Science Foundation under the REAL grant program, award #0816359.

References

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

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the thirty-third annual conference of the cognitive science society* (pp. 2469–2474).

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425. doi:10.1007/s11336-013-9350-4.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Howard, R. A. (1960). *Dynamic programming and markov processes* (1st ed.). Cambridge, MA: MIT Press.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. doi:10.1177/014662168200600301.

LaMar, M. M. (2014). *Models for understanding student thinking using data from complex computerized science tasks (Unpublished doctoral dissertation)*. Berkeley: University of California.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning* (pp. 59–81). New York: Springer.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & Hassabis, D. (2015, February). Human-level control through deep reinforcement learning. *Nature,518*(7540), 529–533.

National Research Council. (2014). *Developing assessments for the next generation science standards*. Washington, DC: The National Academies Press.

Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning* (pp. 663–670) (2000).

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: Wiley.

Rafferty, A. N., LaMar, M. M., & Griffiths, T. L. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, *39*(3), 584–618.

Red Hill Studios. (n.d.). *Lifeboat to mars*. Retrieved from http://www.pbskids.org/lifeboat.
Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144.
Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River: Pearson.
Rust, J. (1994). Structural estimation of Markov decision processes. In R. Engle & D. McFadden (Eds.), *Handbook of econometrics* (pp. 3081–3143). Amsterdam: Elsevier Science.
Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., et al. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling*, *55*(4), 335–360.
Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.
Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest [computer software and manual]*. Camberwell, VIC: Australian Council for Educational Research.

Manuscript Received: 19 APRL 2016
Final Version Received: 30 MAR 2017
Published Online Date: 26 APR 2017