Characterizing Pause Behaviors in a Science Inquiry Task

Caitlin Tenison and Burcu Arslan
(ctenison@ets.org) (barslan@ets.org)
Cognitive and Technology Sciences Center
Educational Testing Service, Princeton, NJ, 08541, USA

Abstract

In inquiry-based learning tasks students are actively involved in learning knowledge and skills through experimentation. The success of these activities largely depends on student's inquiry practices. While traditional assessment infers student competency from their responses and problem-solving steps, the pauses between these actions provide a valuable source of information. Pauses during inquiry tasks capture a wide range of productive and unproductive activities such as planning, reasoning and mind-wandering. We present efforts to characterize the pause behaviors during a science inquiry task using hidden Markov modeling. We explore how theory can inform data driven modeling approaches, describe initial evidence of meaningful pause states, and consider the limitations of this approach for supporting inferences about students' science inquiry practices.

Keywords: Science inquiry, Pauses, Process modeling, Hidden Markov modeling.

Introduction

Several probabilistic and data mining approaches have been used to infer student knowledge and skills from process data (Levy & Mislevy, 2016). Most of these approaches focus on the correctness of the steps that students perform (Yudelson, Koedinger & Gordon, 2013). While effective in domains where correctness is clearly defined, these approaches have limited application in inquiry-based interactive tasks in which students discover and apply relevant knowledge and skills, form and test hypotheses, and then reflect on the outcome of those tests. For these tasks, considering the problem-solving process and the cognitive mechanisms underlying those actions are important when making valid inferences about students' science inquiry practices. While we can draw inferences about the problem-solving process through modeling the actions students take, modeling the pauses between actions can also support inferences about underlying cognitive processes supporting those actions. The goal of our current research is to identify methods for characterizing pauses in the problem-solving process and establish what these pauses contribute to the measurement of inquiry skill. We explore how theory can inform data driven approaches and describe initial evidence, while weighing the limitations of this approach for supporting inferences about students' science inquiry practices.

Evaluating Pauses in Educational Activities

Early studies on pause behaviors during problem solving indicate that pauses between actions provide insight into the processes supporting task completion (e.g., VanLehn, 1991). Given the variability in the number and length of pauses between students, there have been several attempts to use this information to characterize the proficiency of the student (Pelanek, 2014; Dang, Yudelson, & Koedinger, 2017). In an expansion of Bayesian Knowledge Tracing (BKT), Pelanek (2014) incorporates timing into the approach to improve estimates of skill level of students using a tutoring system. In this work longer pauses decrease the probability that the student has mastered the skill. This assumption may be correct when fluency is the goal of learning but is less appropriate when pausing reflects productive behaviors. For example, in their model of diligence, Dang, Yudelson, and Koedinger (2017) propose the use of both time and performance to separate the impact of productive and unproductive pauses on their measure. Other studies take a more unstructured approach and use pauses as input for machine learning algorithms predicting different constructs such as carefulness (Banawan, Andres, & Rodrigo, 2017). A similarity across all these approaches is that they aggregate pauses within the problem-solving process to create general measures of time on task rather than considering the occurrence of these pauses within the process data.

Pauses at different points throughout an educational activity are indicative of different cognitive activities. Prior research using pauses to identify periods of wheel-spinning, gaming the tutor and productive persistence rely on expert qualitative coding and the structure of the tutor environment to identify when a pause is likely to indicate these behaviors (Paquette et al., 2014; Aleven et al., 2004). While these pauses are informative for modeling student behavior, they require human coding and, for this reason, tend to be used to characterize pauses in tutors with a limited space of actions.

Inquiry Learning Activities

Inquiry-based learning environments support students in learning concepts through the exploration and development of general learning behaviors and strategies. Inquiry learning activities are popular in science education as a means of teaching scientific principles through the application of the scientific method. Interactive science simulation environments require students to interact in an open-ended task to generate responses that help support the collection of evidence about what students know (declarative knowledge) and can do (procedural knowledge). Pauses within inquiry learning environments can reflect a blend of productive and off-task behaviors. The nature of these pauses can be inferred by using the process data to understand the context in which the pause occurs. However, within open-ended tasks this type of inference is non-trivial because students can produce a wide range of distinct actions and the underlying cognitive processes of these actions are not directly observable (Ercikan & Pelligrino, 2017).

Prior attempts to model student pause behavior in scientific inquiry activities include both data-driven approaches for identifying behavioral patterns and theory-based approaches that model the reasoning and learning involved in scientific inquiry-based tasks. In their work studying student experimentation with an electrical circuit simulation, Perez and colleagues (2017) used coded log data to compare sequences. They found no difference in overall pause frequency across low and high performing students; however, high performing students paused more than low performing students before and after running experiments. This method requires careful coding of the log data and assumes no measurement error, which can be problematic in situations where there is both variability and the potential for interruptions in strategy execution. Theory based approaches have the flexibility to capture the variation in behavior within a simulation environment. In their Simulated Psychology Lab (SPL) task, Schunn and Anderson (1998), created an environment where people could design research studies, collect simulated data and manipulate that data to draw conclusions. Using the SPL environment Schunn and Anderson compared the actions of human subjects to actions of their cognitive model of scientific discovery. This SPL model was able to capture variation in scientific inquiry behaviors due to the structure of the environment and prior experience. While the SPL model was not used to predict timing data, it provides insight into the activity that occurs during the pauses between actions in a scientific inquiry task.

Current Study

In the current study, we model the pause behaviors of students as they interact with a science simulation task designed to assess students' science inquiry practices in relation to the concept of saturation (i.e., maximum concentration) and control-of-variable strategy (Figure 1). This environment provides students with the tools to run experiments, organize data, and report conclusions. The interactive nature of this environment provides students with considerable variability in how they complete the task(e.g., number of trials, strategy use, timing of actions). This variability while likely capturing meaningful differences in science inquiry ability also makes comparison across students challenging.

We use hidden Markov models (HMMs) as an exploratory technique to distinguish different pausing behaviors as they occur in the context of student actions. HMMs capture the probabilistic transition between latent states in sequential time steps. The strength of these models is the Markovian assumption that the probability of the current state is driven by the previous state and the currently observed behavior. HMMs have been used in the educational data mining community to identify behavioral patterns that can be linked to meaningful cognitive states (e.g., BKT) and strategies (Tenison & MacLellan, 2014). We are not aware of any prior work using HMMs to model pause behaviors within student process data from adaptive learning and assessment environments.

The focus of our modeling effort is to characterize student's pause behavior in a science inquiry task. We hypothesize that the pauses observed during inquiry represent distinct cognitive activities. We further propose that optimal characterization of the processing occurring during pauses will provide useful information to improve our assessment of knowledge skills and abilities of students. We use HMMs to address the challenge of characterizing pauses from individual process data. We use theory to guide our construction of these models but allow data to drive the models we fit. To evaluate the descriptiveness of the model, we use three metrics: 1) sensitivity to group differences, 2) prediction of correct conclusions from patterns of pauses during the scientific inquiry activity, and 3) agreement with validated external measures of scientific inquiry skills.

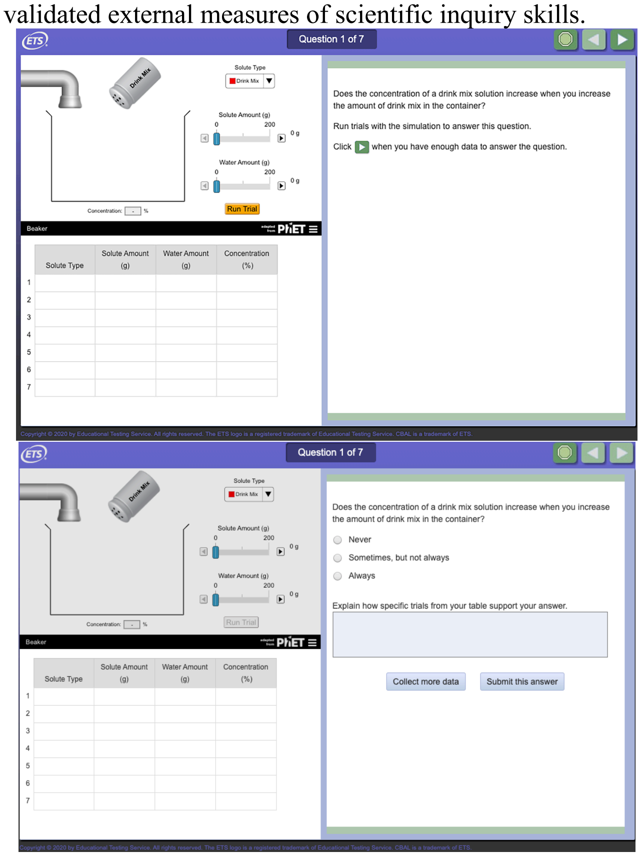

Figure 1. A screenshot of question 1 in the concentration simulation: a) Science inquiry screen, b) Answer screen.

Methods

Participants

Two-hundred-seventy-three students in 6th and 8th grades were tested with a concentration simulation task. Six students were excluded from the analysis due to process data problems. The final data included 134, 6th grade students (N_{Female}= 72) and 131, 8th grade students (N_{Female}= 67). Two students had no grade information associated with their data.

Materials

The concentration simulation task we used was originally developed by PhET (Wieman et al., 2008) and modified for the purpose of the study to include selected and constructed response questions (see Finn, 2018 for a more detailed task description). The simulation was an HTML5 application written in JavaScript and delivered through a standard web browser. In total, there were 7 questions in the task. In the scope of this study, we focus on students' science inquiry behavior (i.e., observable actions and pauses between actions) as students complete the first question within the task, together with, their submitted answers to that question.

This question asks students to run experiments to investigate whether the concentration of a solution increases when the amount of solute increases. Students were instructed to click the next button when they had enough data to answer the question (see Figure 1a). The open-ended nature of the task allowed students to prepare and run experiments by: (i) running as many trials as necessary to give an answer, (ii) setting any value between 0 and 200g for both solute and water amounts, (iii) using different strategies during investigation (e.g., control-of-variable; varying, increasing, or decreasing both variables at the same time). For each student run simulation, results are updated in a 'data-table'. Students can manipulate this data by reordering or deleting experimental records from their data-table. While interacting with the data-table is not required, it is meant as a workspace for students to organize the results of their investigations when drawing conclusions about properties of the concentration solution. After clicking the next button, the answer options for the question together with a constructed response box asking students to justify their selected response appeared on the second screen (see Figure 1b).

Process Data Representation

The representation of the student's scientific inquiry activities that we use to fit our model impacts the descriptiveness of that model. The raw data logfiles the system produced recorded detailed information about specific student actions and system events. From these logs, we identified 10 general categories that capture the actions corresponding to the subcomponents of the concentration simulation task (Table 1). We chose these categories to align with the top-level goal structure of Schunn and Anderson's SPL model (1998). For simplicity, we refer to these categories as '*Actions*' but use labels that indicate whether or not these actions were produced by the student or the interface.

Actions generated by the simulation environment introduce new information to the student. These actions represent standard instructions along with updates to the data collection table based on the experiments that students run. Student actions are changes made to the environment in response to the information the environment provides them. These actions unfold over time. Actions are separated by variable periods where no activity is logged. We refer to these periods as '*Pauses*'. In our analysis, we are interested in characterizing the types of pauses that reflect periods of inactivity that are within the control of the student. We do not analyze the pauses that reflect the time it takes for the simulation environment to run the simulation (e.g., the pauses between the start and end of the animation).

Table 1: Average Percentage (SE) of Session Comprised by Actions types and a Sample Action Sequence.

Representation of Actions	**Percentage**
Experiment Preparation	31.8 (1.2)
Experiment Run	8.6 (.23)
Table Manipulation	4.1 (.5)
Answering Questions	6.7 (.43)
Collect More Data	0.63 (.06)
Done	2.3 (.15)
Interface: Error Message	0.45 (.12)
Interface: Load First Page	2.3 (.15)
Interface: Load Second Page	2.6 (.13)
Interface: Update Table	8.6 (.23)
Sample Action Sequence:	
Interface_Load_FirstPage -> Long_Pause -> Experiment_Prep -> Med_Pause -> Experiment_Prep -> Experiment_Prep -> …	

The action sequence representation (Table 1, bottom row) captures the sequence in which actions occur: however, it does not capture specific temporal information about when these events occur. Prior work using pauses to model cognitive processing and proficiency suggests that the length of the pause is an especially important indicator of what is occurring during that period (Paquette et al., 2014). We explored several different methods of representing the data to capture differences in pause length in HMMs. We considered representing actions as a binned timeseries: however, under this representation pauses dominated the sequences and previous research has indicated HMMs are sensitive to over dispersion (Olteanu & Ridgway, 2012).

Instead, we chose to assign pauses to ordinal categories based on length. Pauses less than 250 ms were ignored because on average motor preparation takes 250 ms (Anderson, 2007) and if an action is preceded by such a pause, it was unlikely to reflect meaningful cognitive activity. For pauses longer than 250 ms, we used the 25th and 75th percentiles as cut-points to categorize pause durations (Figure 2). Pauses between 250 ms and 1.3s were labeled "Short Pauses", pauses between 1.3s and 6.2s seconds were considered "Medium Pauses", and pauses greater than 6.2s seconds were labeled "Long Pauses". In Figure 2, we illustrate the distribution of pauses longer than 250 ms across all sequences with vertical lines indicating category cut-offs. On average, pauses lasted 6.4 s (SD = 12.8). Pauses accounted for 27.4% of the action sequences described in Table 1 (Short: 7.2% (.38), Medium: 16.6% (.45) and Long: 8.2% (.23)).

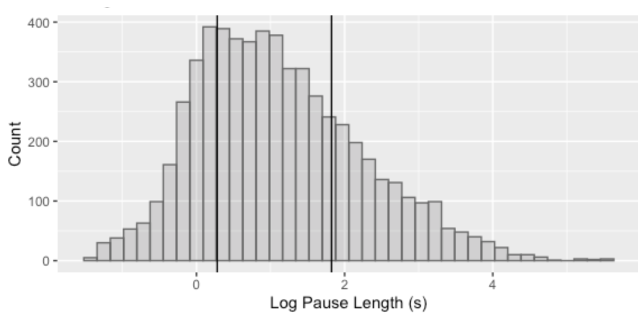

Figure 2: Histogram of the duration of pauses (log-seconds). Vertical lines indicate cut-points.

Hidden Markov Models

Our aim in fitting the HMM to the student's action sequences is to identify hidden states that help us characterize distinct categories of pauses from the context in which they appear and provide us with useful descriptive information about how pauses are expressed in the problem-solving process. We used the R package seqHMM (Helske & Helske, 2017) to fit our HMM models. We use random priors to initialize our emission and transition probabilities, with the exception of the done state which was given a 0 probability of the model starting in that state and a 0 probability of transitioning from that state to any other state. This means the model has no expectation about the types of states present nor any expectation about how people might be moving between these states. For each model we ran the EM algorithm 10 times with randomized starting values for the transition and emission probabilities to avoid fitting local optima.

Results

Models of Pausing Behavior

We fit HMMs with between 3 and 25 states to the data. We used Bayesian Information Criteria (BIC) to determine which model best fit the data while also penalizing for added parameters to avoid overfitting (Figure 3), lower values indicate better model fits. We found that a 16-state model best fit the data (BIC 40504.3, log likelihood -17,6661.5). The next best fitting model (18 state) has a BIC 544.6 points higher than the 16-state model. Typically, a BIC difference greater than 10 is considered strong evidence against the higher BIC value.

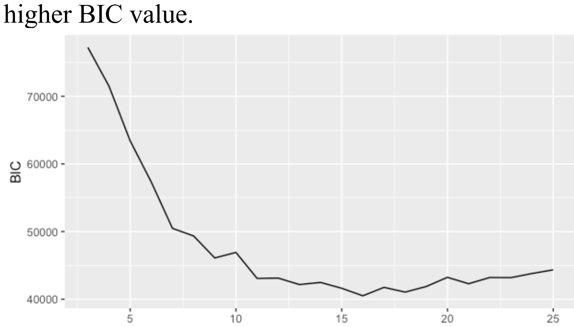

Figure 3: Model fit (BIC) for 3 through 25 state models.

Interpreting Hidden States

We illustrate the structure of the best fitting HMM in Figure 4. The nodes represent the hidden states, the colors of the nodes reflect the probability of that hidden state emitting action events (color coded in the legend of Figure 4). Hidden states emitting pauses show the greatest division across different observable actions. The arrows between nodes represent the transition probabilities, with labels and density reflecting specific probability. For readability we grouped related action states and do not display transition probabilities less than .05.

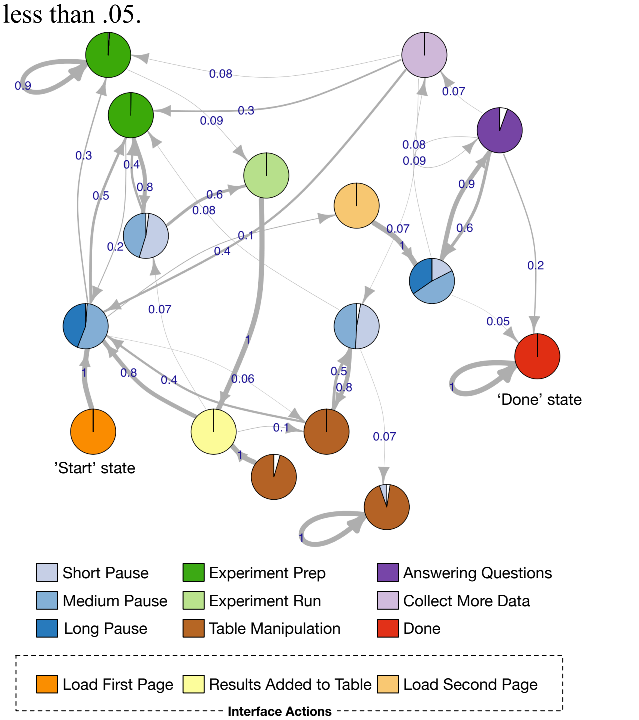

Figure 4: 16-State HMM.

Action states. We can distinguish the states our HMM fits into two categories, *action states* and *pause states*. Our 16-State model fits several states that have a high probability of emitting the same action but fit very distinct transition probability profiles. All of these states had a probability greater than or equal to 95% of emitting to a single state. The only exception, a table-manipulation state, has a 7.6% probability of also emitting a short pause. While these hidden states are likely to emit a single action, the model does make an interesting separation of experiment-preparation and table-manipulation states. For both types of actions the model separates action states that have a high transition probability back into themselves and action states that have a high transition probability into pause states. This distinguishes actions that are swiftly executed as part of a sequence as opposed to actions separated by pauses. This distinction of 'swift' and 'thoughtful' actions could reflect differences in when planning or strategy selection occurs (before or during task execution), or various maladaptive behaviors (e.g. wheel-spinning, mind-wandering).

Pause states. Our primary interest in this work is to explore whether this modeling approach can be used to separate pause states that represent periods of distinct cognitive processing. Our model fit four distinct pause states. We characterize these pauses as they appear in Figure 4 from left to right:

1. **Processing new information and goal setting:** This state captures the pauses that occur immediately after the task starts (96%), and after simulation results are added to the data-collection table (80%). Other states with over a 40% chance of transitioning into this state include a table-manipulation state and the state capturing the decision to run more experiments. From this pause state, students have a high probability of transitioning into an experimental preparation state (75%) or deciding they have collected enough data to progress with the task (14% probability of loading 2nd page). This state consists primarily of medium and long pauses, and an analysis of pause durations of this state indicates an average pause length of 10.4s (SE = 3.9). Given the transition probabilities and length of these pauses, we hypothesize that this state captures an amalgam of processing of new information, deciding to run an experiment or continue to the second page, and planning the experiment.
2. **Experimental Investigation:** This pause state connects to action states involved in preparing and running experiments. These pauses however are relatively short (M = 1.6s, SE = 0.43s). This state may distinguish thoughtful (or aimless) experimental preparation from swift experimental preparation.
3. **Data Manipulation:** This pause state has a high probability of transitioning to (80%) and from (54%) one of the table-manipulation states. As with experiment running this distinguishes thoughtful and swift table-manipulation states. Pauses in this state are relatively short (M = 2.0s, SE = 0.7).
4. **Reflection on Questions:** Our final pause state connects activities on the second part of the science inquiry task (question answering, finishing, and deciding to collect more data). The longer length of these pauses (M = 8.6s, SE = 5.2s) likely capture the time participants spend encoding the answer options, reading and reflecting on responses to the two questions on this page.

Our HMM distinguishes between the scientific inquiry and question answering activities as well as between planning activities, which take longer and capture the switch between high-level task goals and task execution decisions that tend to be much shorter and distinguish thoughtful and swift actions. While this information provides us with a compelling descriptive account of how students complete this inquiry task, we are still limited in our ability to infer the specific cognitive processes that occur during these pause states.

The Role of Pauses in Scientific Inquiry

Our goal in characterizing the different pause states present within the process data of students is to use this information to improve our assessment of scientific inquiry. In our first step to determine if this behavior characterization is informative, we consider 1) whether we see differences in pause behavior between our 6th and 8th grade cohorts, 2) whether the occurrence of different types of pauses are predictive of students answering the question correctly, and 3) whether pauses explain variance in validated measures of scientific inquiry.

We compared the relative proportions of activity spent in the four different pause states of 6th and 8th graders in our sample. We used an independent 2-group Mann-Whitney U test to account for non-normality of the data. Using this test, we found no significant differences in the proportion of activity spent in the four pause states, other than a marginally significant difference in the Data Manipulation Pause state ($W = 9648.5$, $p = .055$), which was more frequently observed in 6th graders compared (1.6%) with 8th graders (.9%).

The scientific inquiry activity we modeled includes a multiple-choice question that asked students to judge whether the concentration of the mixture always increases. This question measures student understanding of the principle of saturation which could result from either their inquiry practices or prior domain knowledge. We tested whether the proportion of the time spent in the different pause states during the inquiry activity were predictive of whether participants got the answer of the multiple-choice question correct. Using a logistic regression, we included all four pause states as factors to predict score on that item. The overall variance explained by this model is low (McFadden pseudo $R^2 = .02$). Two of the pause states were marginally significant: Information Processing (OR: .003, $\beta = -5.8$, $z = -1.7$, $p = .086$) and Experimental Investigation (OR: .25, $\beta = 3.3$, $z = 1.8$, $p = .078$). These low odds ratios indicate students with greater pause activity are more likely to draw the wrong conclusion.

Prior to interacting with the task, students were administered the Waves Benchmark Assessment (WBA) as part of the SimScientists assessment suite developed by WestEd (Quellmalz, Timms, & Buckley, 2010). This measure uses a different science domain but attempts to evaluate the inquiry skills of students. As a first step in determining if the pause states identified by our HMM could provide information about the inquiry skills of the student, we look at the concordance between the pause behaviors and the WBA measure. We fit a linear regression model to measure how much of the variance in the WBA measure we could capture using the proportion of student actions within the four HMM pause states. We started with a maximal model with main effects for all four states and performed stepwise model selection using Akaike information criteria (AIC) to compare fits. Our final model indicated a significant collective effect of the Experimental Investigation, Data Manipulation, and Reflection on Questions pause states ($F(3,188) = 7.6$, $p < .001$, adjusted-$R^2 = .09$, BIC = 1304.6). To test if the HMM states account for more variance in the WBA measure than the raw pause actions, we fit a separate linear regression using the proportion of short, medium, and long pauses in a student's action sequence to predict the WBA measure. Using stepwise AIC model selection, we found that the maximal model best fit the data (F(3,188)=4.8, *p* <.005, adjusted-R^2= .05, BIC=1312.3). This model does not fit the data as well as, nor explain as much of the variance in the WBA score as the pause states model.

Discussion

In this paper, we show how an unsupervised modeling approach can be used to characterize pauses in the problem-solving process and explored what these pauses contribute to the measurement of skill. Our best fitting model identified four distinct pause states and split action states around experimental preparation and table manipulation activities into separate swift and thoughtful action states. These results suggest that pauses capture a range of processes and aggregation across pauses obscures meaningful variation in students' inquiry practices. We found weak evidence that 6th graders and 8th graders pause with similar frequency, but 6th graders pause slightly more when manipulating data. Across grades, students who paused while setting up experiments and in between inquiry activities were score incorrectly on the subsequent multiple-choice item. Finally, we found that pauses around experimental preparation, data manipulation and question answering varied in concordance with student's science inquiry ability, and that our characterization of pauses explained 4% more variance than considering only pause length. The finding that pauses explain a small proportion of variance in scores on the multiple-choice item and the WBA measure illustrates the challenge of using this information to evaluate the inquiry process. The conclusions students draw only partially reflect inquiry ability and future work validating models of inquiry process would benefit from more direct measures of planning, investigation, and analysis skills.

It is unlikely HMMs can discover the structure of problem-solving strategy at the same granularity of cognitive architectures (e.g., ACT-R; Anderson, 2007); however, these models can be used to provide a computational formulation of behavior patterns that balances an adherence to cognitive science theory, parsimony and conformity to data. Our goal in fitting a descriptive model was to better understand differences between individuals and capture meaningful variation in skill. While we examine group differences in a post-hoc analyses, in future research this information could be used as covariates within our HMM to guide model fitting (Helske & Helske, 2017). Such an approach would be especially appropriate in situations where there are clear hypotheses about how factors such as student ability and group membership drive differences in behavior patterns. The results of such models can be used to focus the subtasks we construct cognitive models to capture.

The precision of HMMs in capturing cognitive activity is limited by our ability to observe the thinking of the student. This is especially obvious in our first pause state which combines processing new information, deciding the next action and planning the execution of that action. In future work, we plan on exploring several avenues for improving our ability to distinguish pauses. In the current study, we found that how we represented pauses impacted the descriptiveness of our model. One approach for improving the distinction between pause states is to extend our representation of the actions students take to include more information about the actions. Including information about the data collection strategy students use may help us to identify pause behavior related to specific experiment goal subtypes such as those identified in the SPL model (Schunn & Anderson, 1998). We believe the combination of data and theory within these models will lead to promising avenues for assessing inquiry skills.

Acknowledgments

This work was funded by Educational Testing Service.

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004, August). Toward tutoring help seeking. In *International Conference on Intelligent Tutoring Systems* (pp. 227-239).

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford: Oxford University Press.

Banawan, M. P., Andres, J. L., & Rodrigo, M. T. (2017, October). Predicting student carefulness in an educational game for physics using semi-supervised learning. In *Conference on Information Technology Education* (19-21).

Dang, S., Yudelson, M., & Koedinger, K. R. (2017, April). Detecting diligence with online behaviors on intelligent tutoring systems. In Proc. *ACM Conference on Learning@ Scale* (51-59).

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes*. New York, NY: Routledge.

Finn, B. (2018). Metacognitive evaluations during science simulations: How do ratings of confidence and understanding relate to science assessment inquiry processes? *Archives of Scientific Psychology*, 6(1),117-129.

Helske, S., & Helske, J. (2017). Mixture hidden Markov models for sequence data: The seqHMM package in R. arXiv preprint:1704.00543.

Levy, R., & Mislevy, R. J. (2016). *Bayesian Psychometric Modeling*. Boca Raton, FL: CRC Press.

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014, July). Towards understanding expert coding of student disengagement in online learning. In Proc. *Cog. Sci. Soc.*

Pelánek, R. (2014). Application of time decay functions and the Elo system in student modeling. In Proc. *Educational Data Mining*.

Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., & Roll, I. (2017). Identifying productive inquiry in virtual labs using sequence mining. In Proc. *International Conference on Artificial Intelligence in Education* (287-298).

Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: The

Calipers project. In Proc. *International Journal of Learning Technology*, 5(3), 243-263.

Schunn, C. D. & Anderson, J. R. (1998). Scientific discovery. In J. R. Anderson, & C. Lebiere (Eds.). *The Atomic Components of Thought*, 385-428.

Tenison, C., & MacLellan, C. (2014). Modeling strategy use in an intelligent tutoring system: implications for strategic flexibility. In Proc. *International Conference on Intelligent Tutoring Systems* (pp. 466-475).

Wieman, C., Adams, W., & Perkins, K. (2008). PhET: Simulations that enhance learning. *Science,* 322, 682–683.

VanLehn (1991). Rule acquisition events in the discovery of problem-solving strategies. *Cognitive Science*, 15(1), 1-47.

Yudelson, M., Koedinger, K., Gordon, G. (2013) Individualized Bayesian knowledge tracing models. In Proc. *International Conference on Artificial Intelligence in Education* (pp. 171-180).