Process Matters! Novel Approaches to Using Process Data for Psychometric Research

Ni Bei

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor in Philosophy

University of Washington

2023

Dissertation Committee:

Elizabeth Sanders, Chair

Dagmar Amtmann

Min Li

Deborah McCutchen

Chun Wang

Program Authorized to Offer Degree:

College of Education

University of Washington

ABSTRACT

Process Matters! Novel Approaches to Using Process Data for Psychometric Research

Ni Bei

Chair of the Supervisory Committee:

Elizabeth A. Sanders

College of Education

Logs of keystrokes, clicks, eye tracking, mouse movement, action sequences, and time stamps – known as "process data" – have the potential to provide new insights into how people are thinking about item tasks and ideas as they respond to them. The purpose of this dissertation is to extend the prior research on novel analytic approaches – namely network and latent class modeling – for incorporating action sequence process data in psychometric research, this time within an educational gaming environment. Specifically, I used time-stamped action records for selected items of an online elementary level math game to demonstrate four methods: 1) computing descriptive network statistics on each student's item action sequence for predicting student game performance and engagement for a sample of $N = 20$ fourth grade students that were assessed on a common set of four math modeling items; 2) using inferential network analyses (exponential random graph models) on each student's item actions for predicting student game performance and engagement (same set of 20 students and items); 3) conducting hidden Markov models (HMMs) on latent state action-to-action transitions for $N = 500$ students on the first (entry item) of the math modeling game, and comparing transition patterns by student engagement/performance levels; and 4) using latent class analysis (LCA) to group and predict

student outcomes (same set of students and entry item). Consistent with previous research, results showed that each approach revealed different insights into item properties and how those properties connect with student performance and engagement outcomes. This said, the unique game environment of this assessment dictated that item properties would be different at different phases of the game, such that the connections between network or latent factor properties would differ with student outcomes. Further, because real data was used, some results may have limited generalizability. As such, future research directions should include evaluating these analytic approaches using simulated data with varied sample sizes, numbers of action sequences, and action sequence sparseness levels. Further, it would be helpful to qualitatively investigate best practices in connecting these types of model results with actionable feedback for assessment developers. Despite these limitations, this dissertation was able to successfully demonstrate the potential usefulness of applying network and latent class analysis approaches in analyzing item action sequence data to better understand item and student outcomes within an online math modeling game environment.

Keywords: psychometrics, process data, action sequence, network, hidden Markov model, latent class analysis, math game environment

ACKNOWLEDGEMENTS

I would like to thank all my committee members, as well as all the faculty members and students in the Measurement and Statistics program at the University of Washington's College of Education. It has been an amazing experience, and I am sincerely grateful to be part of this warm family, as well as for their support in academic and daily lives.

In particular, I am very grateful to my committee members for their support in completing my dissertation, which would not have been possible without their guidance and wisdom. First, I would like to thank my beloved advisor, Dr. Liz Sanders, for her deep kindness and generous help in guiding me through my daily studies and research, and ultimately to my doctoral degree. She has always been my role model and has always encouraged me to achieve as much as possible. I would also like to thank Dr. Min Li, who always cared about me and inspired me with new ideas, and Dr. Chun Wang, who always provided insightful feedback on my work. I also thank Drs. Deborah McCutchen and Dagmar Amtmann, who were always kind and supportive, especially during the pandemic. I am also very grateful to Roy Szeto for his assistance with obtaining the *RiddleBooks* data from the UW Center for Game Science. Last but not least, I would like to give special thanks to Dr. Qiwei He, my former internship mentor at ETS, for her continuing mentorship over the year and her brilliant inspiration on my third aim!

I also want to express my gratitude to my family and friends for their constant support during every step of my decision to pursue a doctoral degree. I am truly grateful to my fiancé, Pingyuan (Bay) Zhang, who has always loved and believed in me from the bottom of his heart. He has walked me through all the ups and downs, shared every joyful and difficult moment, and I couldn't be more fortunate to have him in my life. And finally, thanks to Jax, William, and Citrus, our kitties and "stress relievers" who did a terrific job of keeping me company!

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1.

Process Data in Psychometric Research

There has been a widespread shift from paper-and-pencil assessments and surveys to those that are computerized (e.g., Clarke-Midura & Dede, 2010; Scrimgeour & Huang, 2022), allowing for increased opportunities to collect moment-to-moment information never before available (Bergner & von Davier, 2019). Logs of keystrokes, clicks, eye tracking, mouse movement, action sequences, and time stamps – known as "process data" – have the potential to provide new insights into how people are thinking about item tasks and ideas as they respond to them (Bergner & von Davier, 2019). In one paper, Zhu et al. (2016) used data visualization of NAEP item action-to-action sequences (as individual networks) to group test-takers into different levels of problem-solving efficiency and found meaningful differences between groups on item network descriptive statistics. In another example, Xie and colleagues (2019a) examined student keystroke data and found it was useful in detecting confusing item instructions; by discovering and revising the problematic instructions, the validity of the assessment was improved. And, in yet another example, He et al. (2021) computed the distance between an efficient action sequence for item problem-solving (using expert input) and test-takers' own individual action sequences. Not surprisingly, students who used more optimal action sequences had higher test performance.

The purpose of this dissertation is to extend the prior research on novel analytic approaches for incorporating action sequence process data in psychometric research, this time within an educational gaming environment (as distinct from typical test-taking scenarios). Specifically, I used time-stamped action records for selected items of an online elementary level math game to demonstrate four methods: 1) computing **descriptive network** statistics on each

student's item action sequence for predicting student game performance and engagement; 2) using **inferential network** analyses (exponential random graph models) on each student's item actions for predicting student game performance and engagement; 3) conducting **hidden Markov models** on latent state action-to-action transitions, and comparing transition patterns by student engagement/performance levels; and 4) using latent class modeling to **extract latent classes** (clusters) of students based on their action sequence patterns, and using student engagement/performance to predict clusters. Below I provide a brief overview of research to date on process data use in psychometrics, and then I highlight each of the analytic methods proposed.

Psychometric Research using Process Data

Educational assessments should provide valid and reliable scores to support the adequacy and appropriateness of score inferences. Advances in the development of digital platforms have enabled streaming data collection, thus providing a potentially rich source of information in understanding how students solve a problem rather than merely how well students solve the problem (Bergner & von Davier, 2019; Zhu et.al., 2016). Although the validity of electronic assessments can be complicated, one of the benefits of substituting paper-and-pencil tests with digital testing has been that test-taker actions during the testing process can now be recorded for evaluative use. Researchers have explored various aspects of analyzing and modeling process data, such as machine learning techniques (e.g., Bergner et al., 2014) and event history analysis (e.g., Chen et al., 2019). More specifically, within the psychometrics context, Shu et al. (2017) adapted IRT models in extracting student response features; Xiao et al. (2021) adapted hidden Markov models in analyzing student action sequence clustering and transition at the latent stage;

and most recently, Chen et al. (2022) adapted latent space model within network analysis framework in measuring student performance through comparisons of action networks.

Response time data has also been of interest to assessment researchers, particularly for use in increasing score validity (Dutilh et al., 2019). For example, response time has been widely used in detecting examinee's rapid guessing, item omitting, and disengaging behaviors (e.g., Lu et al., 2020; Sahin & Colvin, 2020; Ulitzsch, von Davier & Pohl, 2020), and in improving measurement precision and proficiency estimations by incorporating response time with existing models (e.g., van der Linden, 2007; Reis Costa et al., 2021a). In addition, other works have investigated the relationships between response time and item performance (e.g., Goldhammer & Klein Entink, 2011) as well as students' response-item interactions and problem-solving strategies (e.g., Goldhammer and Zehner, 2017).

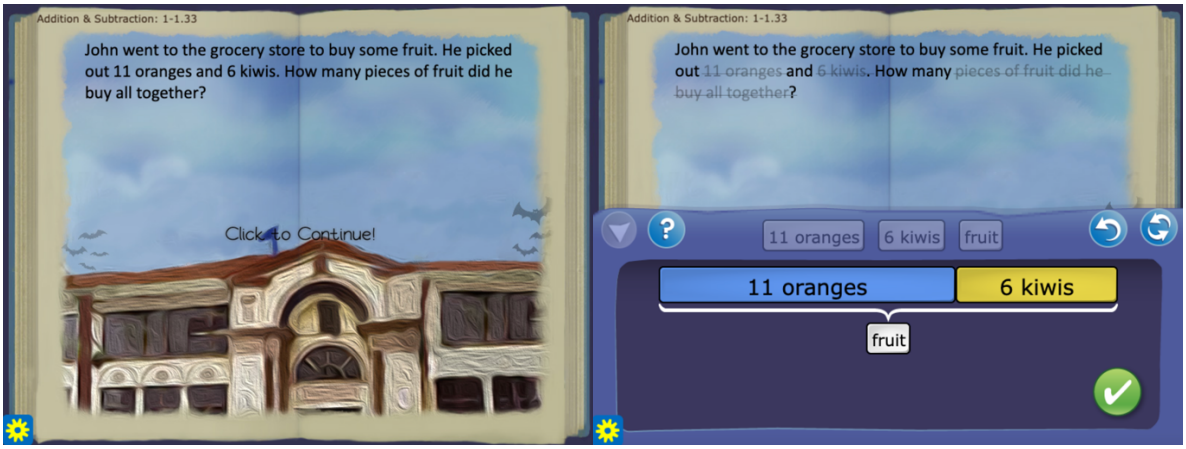
Overarching Research Questions

The core of this dissertation is to demonstrate four analytic approaches to incorporating action sequence process data for psychometrics research, within an educational game environment. The data for each demonstration is from a unique online mathematics game developed by University of Washington researchers called *RiddleBooks* (teacher's release version; Center for Game Science, 2015). The game is geared for elementary and middle school students to measure math problem-solving skills, with student players' action sequences and timestamps (e.g., dragging pieces of information to build a bar model) tracked and collected by a real-time user console on a digital gaming platform. Specifically, the game uses math word problems (WP) in measuring student's math modeling skills through items with story contexts from a variety of themes, which provide students with practical real-world problem situations. WP requires the connection between linguistic and mathematical comprehension, and the success

in mathematics modeling reflects meaningful learning and understanding of concepts (Daroczy et al., 2015; Debrenti, 2015). Figure 1 shows an example item environment and its solution.

Figure 1

RiddleBooks Example Item Environment and Bar Model (Item 511)



To solve the item, a student player needs to use clicks to drag and put together the information pieces of "11 oranges" and "6 kiwis" as two bars, then drag and place the information piece of "pieces of fruit did he buy all together" as a representation of summation of the two bars. After building the bar model, a student player can use the "click" button to validate the model in order to check whether the model is correct or not. The game items consist of different bar models that reflect knowledge of different relationships among numbers, such as part-whole and comparison relationships (Ng & Lee, 2009). During the problem-solving process, student players can also use tools to help them solve the item, such as hint, go back, and redo. In order to investigate student player's math modeling skills as "naive" patterns without pre-knowledge, the focus was on students' first attempts on items (i.e., actions before the first model validation) without hints.

Using item data from the *RiddleBooks* game, I propose to demonstrate four types of action sequence data analyses that can be used for drawing inferences item characteristics or student outcomes, including: action network descriptive statistics and data visualization, action

network exponential random graph modeling (ERGM), hidden Markov modeling (HMM), and latent class analysis (LCA). To further clarify, the first two approaches belong to the domain of network analysis, with the Aim 1 focusing on extracting descriptive network statistics out of student action networks to predict student outcomes and Aim 2 focusing on predicting how action network activities arise given student characteristics through inferential network models; the latter two approaches belong to the domain of latent class analysis, with Aim 3 focusing on extracting latent states and latent state transitions of observed action sequences and Aim 4 focusing on classifying students by their behavioral features in the presence of observed action-step combinations (without transitions). The overarching research questions are as follows.

1) To what extent can these different analytic approaches using action sequence data allow for meaningful inferences about student outcomes, such as performance and game engagement?
2) To what extent can these different analytic approaches using action sequence data allow us to improve assessment validity?

CHAPTER 2.

Aim 1: Describing Item Action Sequence Data as a Network for Prediction

Network analysis focuses on relationships among individuals as well as how those relationships influence individual or group behaviors and attitudes (e.g., Lazega & Snijders, 2016). Born out of sociological theory, network models have been most used to quantify relationships among people. Examples include predicting the spread of disease and drug use, physical and emotional health outcomes, political party affiliations, trade patterns, kinship, and social media memes, to name a few (e.g., Granovetter, 1973; Nasrinpour et al., 2016; Valente et al., 2004; White, 2014). In education, network analyses have been used to understand connections among teachers within and across schools, particularly during or after professional development in the hopes of increasing information sharing amongst participants, thereby improving instruction or leadership (e.g., Thompson et al., 2019; Windschitl et al., 2012).

Although researchers from ETS have analyzed test item action sequence data for making inferences about examinee characteristics, rather than item properties (e.g., He et al., 2021; He et al., 2022; Liao et al., 2020; Xiao et al., 2021), this work has primarily focused on clustering analysis types of approaches. To date, only three studies have hinted at the potential usefulness of network models with item action data. The first was from Zhu et al. (2016) who used data visualization of NAEP test item action sequence networks to categorize learners into different levels of problem-solving "efficiency." The second by Jiao et al. (2021), reiterated this idea as a potentially useful analysis under a range of analytic choices for process data. And most recently, Chen et al. (2022) studied how latent space models (a type of network analysis) could be used to identify latent action positions with PISA test item action data, which was then essentially used to predict student performance.

In short, the prior research on item action sequence data has mostly focused on clustering methods that can derive optimal or "efficient" problem-solving sequences that in turn are used to predict correct *test* item response. In contrast, the present aim focused on summarizing item action *network properties as a whole* (similar to Zhu et al., 2016, and Jiao et al., 2021) and exploring how these network item properties, when combined with student characteristics, may or may not predict item response *as well as post-item engagement in an educational game environment*. In addition, the subset of items examined were those that were administered sequentially, beginning with the first item ever administered; as such, we can see how different network features may or may not predict outcomes as students progress through the game.

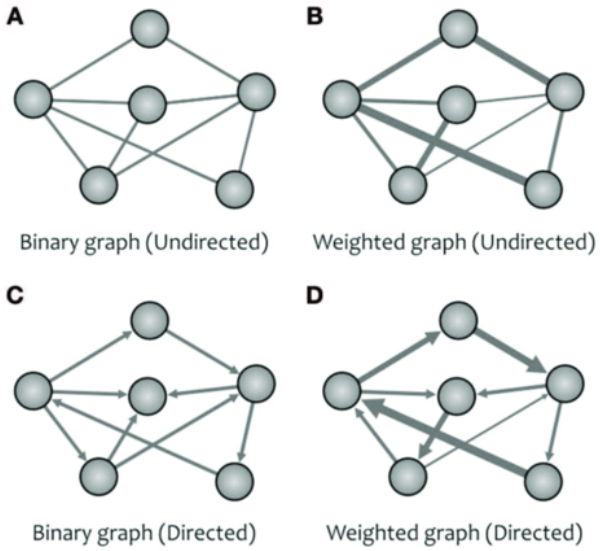
Network Vocabulary

Before going further, it is wise to review network vocabulary. A **network** is a group of entities (often people, but they can also be ideas, and in the case of this dissertation, actions); these entities are called **nodes** (or actors), and the relations among nodes are called **ties**, **edges**, **connections**, or **links**. Between any given pair of nodes (known as dyads), a network can be **directed** or **non-directed**: a directed network consists of ties with directional relations, while a non-directed network consists of ties that are purely associative. For example, if we observe student interactions within a classroom, a directed network could be how students nominate others as their friends, and a non-directed network could be how students talking to each other during a group project. A network can also be distinguished by the values attached to its ties, which makes it **binary** or **weighted**. A binary network only captures the presence of a tie, with 1 indicating presence and 0 indicating absence, while a weighted network captures not only the presence or absence but also the intensity or magnitude of the ties between nodes. For example, we may use a binary network when we are observing whether two students talked to each other

during class, and a weighted network when we are observing how many times the students talked to each other during class. Figure 2 shows the distinctions among four types of networks.

Figure 2

Example Social Networks



Note. Reprinted from "Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: A Systematic Review", by Farahani, F. V., Karwowski, W., & Lighthall, N. R. (2019), *Frontiers in Neuroscience*, *13*, 585, p.7. Copyright 2019 by Frontiers in Neuroscience.

Beyond the basic types of networks, there are also a variety of metrics that can be used to measure linkage patterns and their variability (e.g., Carolan, 2014; Harris, 2014). Network **tie count** is the total number of ties a network has, and network **density** measures the overall connectedness of a network by taking the number of observed ties out of the number of all possible ties in the network (ranges from 0 to 1). A density close to 0 is considered "sparse" and a density close to 1 is considered "saturated". Within a directed network, we can also consider **mutual** tie count and **non-mutual** tie count, which denote the number of mutually connected ties and the number of unidirectional ties, respectively. **Reciprocity** (also known as mutuality) is the percentage of ties in a network where both members in a dyad (i.e., a pair of nodes) send and

receive nominations from and to each other, relative to the number of ties observed in the network.

Other popular network descriptive statistics are the measures of **centrality** or **centralization** in describing network connectivity. In this study, I consider **degree centrality**, **betweenness centrality**, **distance** and **diameter** in describing action networks, within the framework of binary, directed networks. Specifically, **degree centrality** captures the count of ties each node has, and can be divided into **in-degree** and **out-degree** centrality. The former represents the number of ties sent to a node and the latter represents the number of ties received by a node. In-degree centrality is also referred to as "popularity" of nomination, and the most central node is the one with the most ties. Betweenness centrality captures the number of unique shortest paths that a node may be on, serving as a bridge between the flows of other nodes in the network. If a network has a high degree of betweenness centrality, then it indicates that the network has many nodes serving as bridges between other nodes. Lastly, distance captures the average length of a path from one node to another. In this study, only mean distances of networks (i.e., the average distance of all ties in the network) are considered, where a large mean distance indicates more ties in the network. On the other hand, diameter is often referred to as the "longest-shortest path", i.e., the shortest distance between the two most distant nodes in the network. Smaller diameters indicate that fewer steps are needed to travel between nodes, and therefore, that information can travel more efficiently through the network.

Research Questions

The present aim summarizes four unique items' network properties and then explores their utility in predicting student item response performance and continued game engagement. More specifically, action sequence were tabulated and response time data collected from

RiddleBooks in representing action sequences as networks through descriptive network statistics and visualizations using *R igraph* package (Csardi & Nepusz, 2006). In particular, *binary* and *directed* networks are created for each student with a starting action (i.e., starting item) and an ending action (i.e., validating model) when answering each item. The research questions are:

1. How can action sequences be transformed to action networks and represent them effectively using descriptive network statistics and graph visualizations?
2. How can action patterns be used of predicting student performance and engagement in the game?

Specifically, I propose a 2-stage approach: in the first stage, I extracted network-level statistics for each student by item, and in the second stage, I incorporated item-student action network characteristics as predictors in linear and logistic regression models to predict student responses and game engagement (i.e., the count of items attempted after each item).

Method

Data Sources

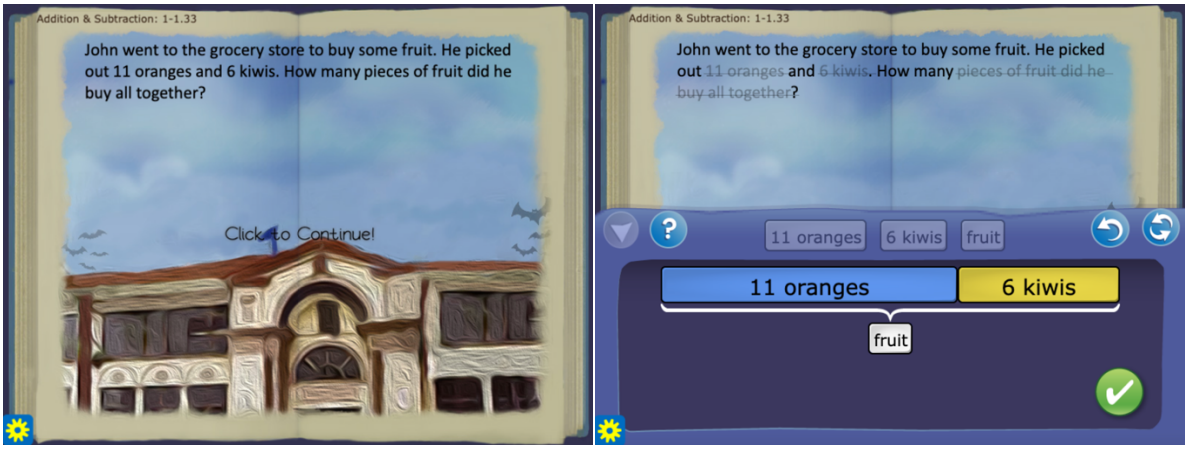
To maximize the sample size for this study, I initially sampled all fourth grade students who had attempted at least five common items from the teacher-supervised version of the game (there is also a version of game that can be used anonymously). This yielded $N = 44$ students. I then removed any students who used fewer than two actions (i.e., start and end), and selected four items with the maximum number of common users. This yielded an analytic sample of $N = 20$ common student players across four items (i.e., Items 511 (start), 701, 961, and 1062). Specifically, *Item 511* was designed for grade level 3-4 with a story theme of "People", *Item 701* was designed for grade level 3-4 with a story theme of "Animal", *Item 961* was designed for grade level 2 and under with a story theme of "Animal", and *Item 1062* was designed for grade

level 2 and under with a story theme of "People". In addition, each item requires a different bar model type to be built for solution, including addition and subtraction models (i.e., part-whole models) as well as multiplication models. Item details are provided below, in the order children received them. Last but not least, I limited the action data to first attempts to obtain "naive" item action patterns that were not confounded by that particular item's practice effects.

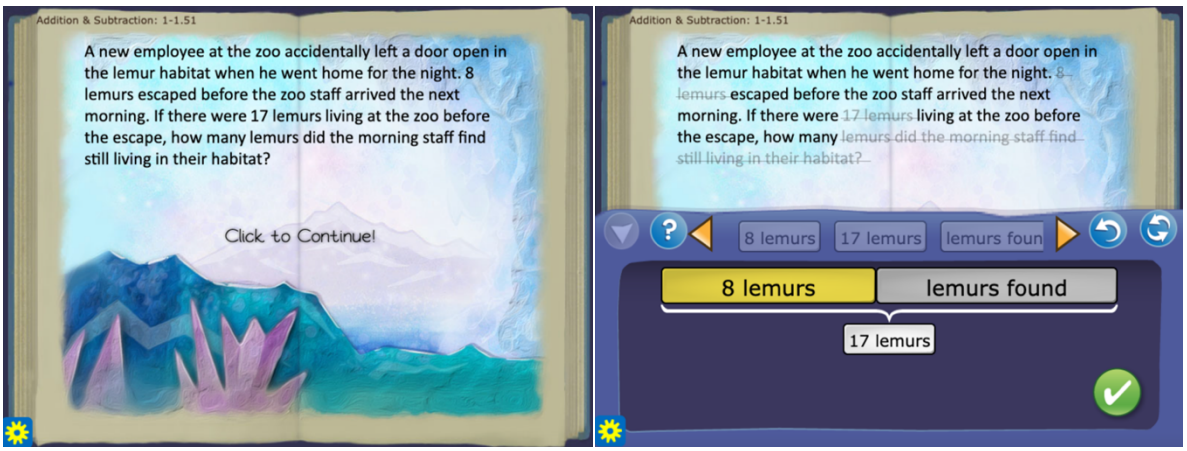
Figure 3

Details about Items Selected for Investigation

Panel A: Item 511 (grade 3-4, bar model type =1a, theme = people, word count = 28)

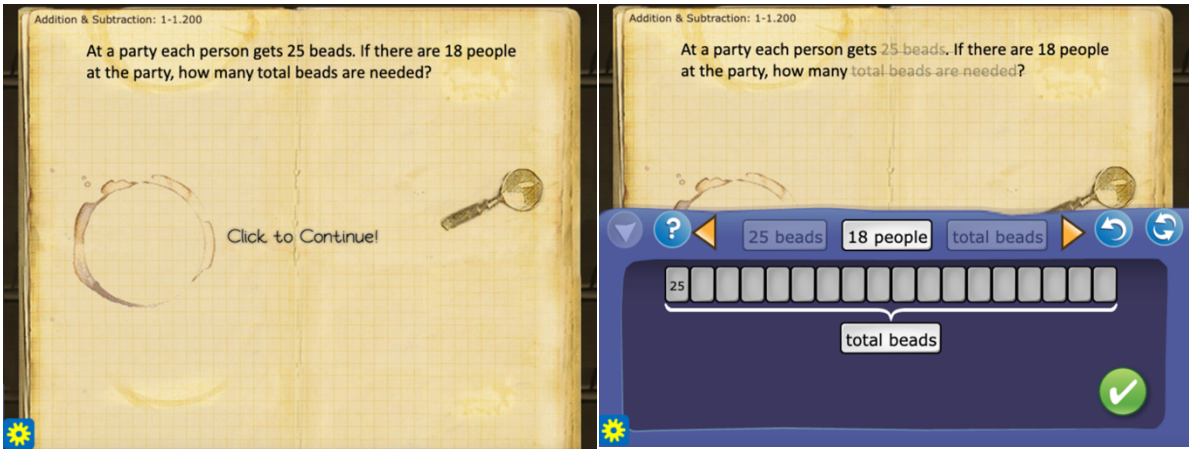


Panel B: Item 701 (grade 3-4, bar model type = 1b, theme = animal, word count = 58)



Panel C: Item 961 (grade 2 and under, bar model type = 2b, theme = animal, word count = 29)

Panel D: Item 1062 (grade 2 and under, bar model type = 3a, theme = people, word count = 22)



Measures

Action Sequences and Networks

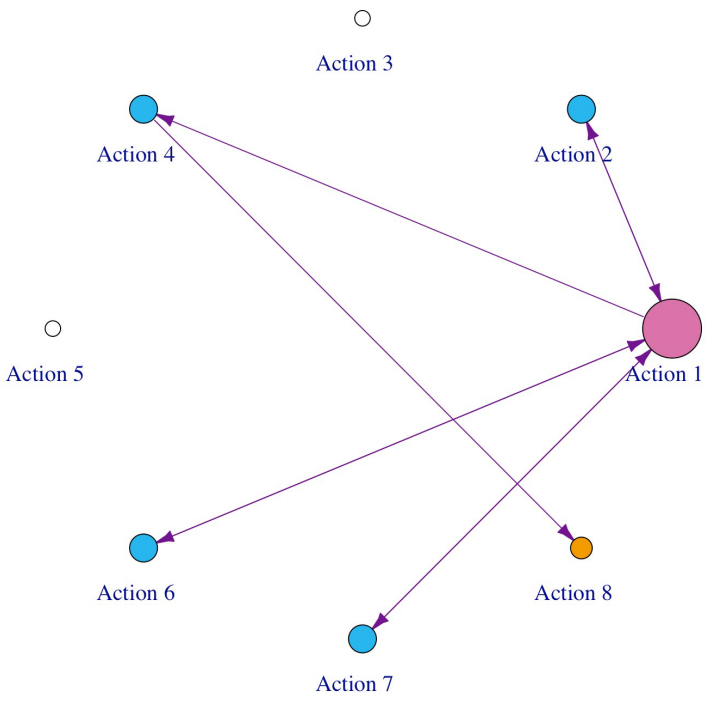
For each item, each student player's raw time-stamped action sequences of all action patterns possible were mapped into action-by-action directed adjacency matrices. Specifically, a total of 11 actions were used on Item 511 and Item 961, thus an 11-by-11 action matrix was created for each user attempting these two items; and a total of 13 actions were used on Item 701 and 1062, thus a 13-by-13 action matrix was created each user attempting these two items. Details of action interpretations and coding book can be found in Appendix A. Figure 3 shows an

example of mapping a given action sequence to a network: in the network graph, a circle represents an action node, and an arrow represents the transition path(s) between action nodes, with a single-headed arrow representing a one-way connection/tie while a double-headed arrow representing a mutual connection/tie between action nodes. Counts averaged 7.30 ties (*SD* = 3.26), *M* = 7.30 ties (*SD* = 5.23), *M* = 4.15 ties (*SD* = 1.87), and *M* = 7.10 ties (*SD* = 3.52), for Items 511, 701, 961, and 1062, respectively. In addition to tie counts, eight other network characteristics were tabulated, including: **non-mutual ties, density, reciprocity, diameter, degree-in, degree-out, betweenness, and mean distance**.

Figure 4

Example Action Sequence Transformed to Action-by-Action Network

Action ID
1
1
2
1
1
1
1
6
1
2
1
7
1
1
4
8



Student Characteristics

Total and average time information was calculated based on the raw timestamps associated with action sequences. I included two response time predictors in the model: 1) student players' average time spent across the entire game, 2) student player total time spent on each item. Results showed that student players spent an average time of 56.04 seconds per item (*SD* = 22.73) across the entire game, and an average time of 35.60 seconds (*SD* = 26.07), 72.86 seconds (*SD* = 59.02), 36.45 seconds (*SD* = 19.17), and 79.26 seconds (*SD* = 47.82) on Items 511, 701, 961, and 1062, respectively.

Overall Student Ability. Students' item performance measured was percent of items correctly answered on the first attempt was also included as a potential indicator of student math modeling ability. The mean was 0.68 (*SD* = 0.13) across the 20 student players.

Pre-Item Practice Effects. A practice effect – or more simply, task familiarity – occurs when an individual is repeatedly exposed to the same stimuli, which can alter task performance in either a positive direction (more efficient problem solving and faster reaction time) or a negative direction (boredom and disengagement) (e.g., Goldberg et al., 2015). In the present study, student players were administered other items after the first item in a computer-adaptive fashion. More importantly, the items after the starting item (Item511) were "unsupervised" in their ordering (i.e., children did not receive the same items in the same order). Given that student performance and engagement on items after the first item will likely depend on prior item experiences, I incorporated 'game practice' as the count of items attempted before the focal item. For Items 701, 961, and 1062, the average previous number of items attempted were *M* = 3.15 items (*SD* = 1.27), 6.55 items (*SD* = 1.43), and 10.85 items (*SD* = 7.05), respectively. This said, these numbers do not represent the total number of attempts, only the number of items attempted.

In fact, students cannot access another item from the computer algorithm until they complete a current item correctly, or stop and restart the game.

Student Outcomes

Item Performance. For each item, I captured each student player's item response in "model validation" after building a bar model in their first attempts, with 1 representing success in building the correct bar model and 0 representing failure to build the correct bar model. Results showed that the averaged percent correct for Items 511, 701, 961, and 1062 were 0.65 (*SD* = 0.49), 0.20 (*SD* = 0.41), 0.85 (*SD* = 0.37), and 0.25 (*SD* = 0.44), respectively.

Post-Item Game Engagement. Researchers have reported positive effects of game-based learning on student achievement, such as enriching learning experience and evoking learner engagement (e.g., Vankúš, 2021). In addition, learning strategies and goals have been found to be different for students with high vs. low self-regulation (Hwang et al., 2021). Because I utilized data from student players in a self-directed online game rather than a required assessment, it is crucial to consider game engagement as a learning outcome. For each user, I defined game engagement as the count of items attempted after each item. Descriptive statistics showed that the average number of items attempted post-item were *M* = 95.90 (*SD* = 61.25), *M* = 92.75 (*SD* = 61.13), *M* = 89.35 (*SD* = 61.08), and *M* = 85.05 (*SD* = 60.98) for Items 511, 701, 961, and 1062, respectively.

Modeling Approaches

I employed a two-stage approach for evaluating the extent to which each of the four items action sequence network information was predictive of student performance on each item, as well post-item game engagement. In the first stage, descriptive network statistics were calculated using the *R* igraph package (Csardi & Nepusz, 2006). In the second stage, I used the *R* MASS

package (Venables & Ripley, 2002) to employ stepwise multiple regressions for each item separately (logistic regression was used to model item performance and linear regression was used to model post-item engagement). Stepwise regression is an exploratory, algorithmic *predictor entry* approach that is useful when there are large numbers of competing predictors that are often quite correlated with each other. Predictors included both item and student characteristics, and for ease of results interpretation: 1) student mean ability across game and network predictors were standardized into z-scores; 2) since time predictors and item counts before and after each item were skewed, I transformed them as the natural logarithm of the values + 1 (**logTime**) and + 1 (**logItemCount**), respectively. Last, I note that for Items 701 and 1062, I did not include the mean diameter predictor due to its exceedingly high correlation with other predictors, coupled with its low correlation with any of the outcomes.

Results

Descriptive Statistics

Network Descriptives. Table 1 displays descriptive statistics for item action-to-action network indices for each item. As can be seen, Items 511 and 701 had the most tie counts (i.e., students did the most exploration for these items) while Item 961 had the least tie counts. Item 961 had the least degrees-in and degree-out, which also confirmed that it had the least number of ties among actions compared with the other items. Additionally, Item 511 had the highest density of 7% while the other items had lower densities between 4% to 5%. Items 511 and 961 had more mutually connected ties compared with Items 701 and 1062. Moreover, Item 511 had the largest centrality of betweenness, mean distance, and diameter, suggesting that the action networks for Item 511 are more complex compared with the action networks of other items (i.e., more "influential" actions that were repeatedly used by students). These indicate that student players

may have performed more exploratory actions with more back-and-forth ties among actions when attempting the starting item (Item 511).

Table 1

Descriptive Network Statistics across Items

Variable	Item 511		Item 701		Item 961		Item 1062	
	M	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
1. TieCount	7.30	(3.26)	7.30	(5.23)	4.15	(1.87)	7.10	(3.52)
2. Non-MutualTie	3.90	(3.14)	6.00	(3.85)	3.85	(1.42)	5.00	(2.51)
3. Density	0.07	(0.03)	0.05	(0.03)	0.04	(0.02)	0.05	(0.02)
4. Reciprocity	0.51	(0.21)	0.14	(0.19)	0.09	(0.24)	0.29	(0.30)
5. Diameter	4.00	(1.45)	3.35	(1.09)	2.95	(1.05)	3.30	(1.22)
6. Degree-In	0.12	(0.05)	0.11	(0.07)	0.09	(0.04)	0.12	(0.05)
7. Degree-Out	0.19	(0.05)	0.12	(0.07)	0.09	(0.05)	0.13	(0.05)
8. Betweenness	0.08	(0.05)	0.05	(0.05)	0.03	(0.03)	0.05	(0.04)
9. Mean Distance	1.96	(0.44)	1.84	(0.41)	1.70	(0.44)	1.82	(0.48)

Note. Values calculated from $N = 20$ fourth-grade students' binary directed networks.

Other Item and Student Descriptives. Tables 2-5 display all descriptive statistics and variable correlations by item. As can be seen, the items were actually set up in a sequence themselves: Item 511 was the first item, Item 701 only had about three previous items, and Items 961 and 1062 had 7 and 11 items before them, respectively. The items varied in their difficulty levels, going from moderately easy to relatively difficult, and back to easy and then difficult again. As will be readily seen, the only consistent pattern is that student's mean ability was significantly correlated with the increase of post-item count (i.e., game engagement) for all items. Other patterns in which indicators were predictive of the outcomes were item-specific, which indicates that each item's particular characteristics were unique. For network statistics, degree-in and degree-out centralities were significantly and positively correlated with a correct item response for Item 511, but not for any of the other items. Reciprocity was only significantly and negatively correlated with a post-item engagement for Item 701, while mean distance centrality was significantly and negatively correlated with a post-item engagement for Item

1062. For other student predictors, pre-item count was significantly correlated with a correct item response for Item 701, but not for any of the other items.

Table 2

Descriptive Statistics and Correlations for the First Item, Item 511

Variable	*M*	*(SD)*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
1. Item Response (1=corr)	0.65	(0.49)	--												
2. Items Post-Item (Log)	95.90	(61.25)	-.12	--											
3. Mean RT (Log)	56.04	(22.73)	.08	.25	--										
4. Mean Ability (Z)	0.68	(0.13)	.25	.50 *	.37	--									
5. Item RT (Log)	35.60	(26.07)	-.03	-.01	.56 *	.09	--								
6. Tie Count (Z)	7.30	(3.26)	.23	-.37	-.05	-.03	.51 *	--							
7. Non-Mutual Tie (Z)	3.90	(3.14)	.04	-.23	-.06	-.08	.45 *	.90 ***	--						
8. Density (Z)	0.07	(0.03)	.23	-.37	-.05	-.03	.51 *	1.00 ***	.90 ***	--					
9. Reciprocity (Z)	0.51	(0.21)	.18	.05	.03	.25	-.31	-.50 *	-.82 ***	-.50 *	--				
10. Diameter (Z)	4.00	(1.45)	-.15	-.15	.15	-.24	.69 ***	.47 *	.55 *	.47 *	-.63 **	--			
11. Degree-In (Z)	0.12	(0.05)	.56 *	-.36	-.08	.28	.01	.54 *	.23	.54 *	.34	-.34	--		
12. Degree-Out (Z)	0.19	(0.05)	.53 *	.09	-.13	.33	-.40	.00	-.15	.00	.40	-.62 **	.59 **	--	
13. Betwness (Z)	0.08	(0.05)	.17	-.26	.01	.06	.48 *	.88 ***	.90 ***	.88 ***	-.64 **	.56 *	.35	.11	--
14. Mean Distance (Z)	1.96	(0.44)	-.03	-.17	.15	-.24	.66 **	.52 *	.62 **	.52 *	-.69 **	.98 ***	-.29	-.51 *	.63 **

Note. $N = 20$ fourth-grade students. Item 511 is the first attempted item in game, therefore there is no “pre” item count.

Table 3

Descriptive Statistics and Correlations for the Second Focal Item, Item 701

Variable	*M*	*(SD)*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Item Response (1=corr)	0.20	(0.41)	--													
2. Items Post-Item (Log)	92.75	(61.13)	.18	--												
3. Items Pre-Item (Log)	3.15	(1.27)	.54 *	.22	--											
4. Mean RT (Log)	56.04	(22.73)	-.05	.25	.08	--										
5. Mean Ability (Z)	0.68	(0.13)	.33	.50 *	.56 **	.37	--									
6. Item RT (Log)	72.86	(59.02)	-.23	.34	.12	.63 **	.11	--								
7. Tie Count (Z)	7.30	(5.23)	-.30	.34	-.30	.32	-.14	.65 **	--							
8. Non-Mutual Tie (Z)	6.00	(3.85)	-.23	.41	-.26	.29	-.07	.66 **	.96 ***	--						
9. Density (Z)	0.05	(0.03)	-.30	.34	-.30	.32	-.14	.65 **	1.00 ***	.96 ***	--					
10. Reciprocity (Z)	0.14	(0.19)	-.38	-.46 *	-.38	-.08	-.54 *	-.06	.28	.08	.28	--				
11. Diameter (Z)	3.35	(1.09)	-.05	.30	.30	.07	.15	.49 *	.50 *	.59 **	.50 *	-.25	--			
12. Degree-In (Z)	0.11	(0.07)	-.23	.41	-.38	.38	-.02	.56 *	.93 ***	.88 ***	.93 ***	.26	.36	--		
13. Degree-Out (Z)	0.12	(0.07)	-.32	.17	-.38	.31	-.26	.56 *	.93 ***	.87 ***	.93 ***	.41	.34	.90 ***	--	
14. Betwness (Z)	0.05	(0.05)	-.31	.29	-.33	.23	-.05	.58 **	.93 ***	.93 ***	.93 ***	.18	.60 **	.86 ***	.83 ***	--
15. Mean Distance (Z)	1.84	(0.41)	-.17	.35	.15	.06	.08	.54 *	.56 *	.66 **	.56 *	-.22	.95 ***	.41	.38	.69 **

Note. $N = 20$ fourth-grade students.

Table 4

Descriptive Statistics and Correlations for the Third Focal Item, Item 961

Variable	*M*	*(SD)*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Item Response (1=corr)	0.85	(0.37)	--													
2. Items Post-Item (Log)	89.35	(61.08)	-.24	--												
3. Items Pre-Item (Log)	6.55	(1.43)	.05	.25	--											
4. Mean RT (Log)	56.04	(22.73)	.06	.24	-.07	--										
5. Mean Ability (Z)	0.68	(0.13)	.16	.51 *	.49 *	.37	--									
6. Item RT (Log)	36.45	(19.17)	.17	-.30	.12	.22	-.01	--								
7. Tie Count (Z)	4.15	(1.87)	-.12	-.02	.13	-.17	.21	.37	--							
8. Non-Mutual Tie (Z)	3.85	(1.42)	-.05	-.01	.05	-.18	.11	.38	.94 ***	--						
9. Density (Z)	0.04	(0.02)	-.12	-.02	.13	-.17	.21	.37	1.00 ***	.94 ***	--					
10. Reciprocity (Z)	0.09	(0.24)	-.01	.10	.32	-.02	.05	.21	.05	-.10	.05	--				
11. Diameter (Z)	2.95	(1.05)	-.02	.29	.01	.05	.44	.17	.59 **	.63 **	.59 **	-.28	--			
12. Degree-In (Z)	0.09	(0.04)	-.04	-.29	.03	-.20	.04	.07	.60 **	.38	.60 **	.30	.03	--		
13. Degree-Out (Z)	0.09	(0.05)	-.26	-.16	.12	-.16	.16	.02	.63 **	.39	.63 **	.32	.12	.92 ***	--	
14. Betwness (Z)	0.03	(0.03)	-.20	-.01	.03	-.12	.30	.16	.86 ***	.74 ***	.86 ***	.08	.67 **	.67 **	.79 ***	--
15. Mean Distance (Z)	1.70	(0.44)	.03	.26	-.01	.13	.43	.25	.60 **	.69 ***	.60 **	-.27	.94 ***	-.03	.05	.61 **

Note. $N = 20$ fourth-grade students.

Table 5

Descriptive Statistics and Correlations for the Fourth Focal Item, Item 1062

Variable	*M*	*(SD)*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Item Response (1=corr)	0.25	(0.44)	--													
2. Items Post-Item (Log)	85.05	(60.98)	-.06	--												
3. Items Pre-Item (Log)	10.85	(7.05)	.35	.06	--											
4. Mean RT (Log)	56.04	(22.73)	.15	.23	-.22	--										
5. Mean Ability (Z)	0.68	(0.13)	.09	.51 *	.32	.37	--									
6. Item RT (Log)	79.26	(47.82)	.01	.20	-.32	.67 **	.39	--								
7. Tie Count (Z)	7.10	(3.52)	-.05	.12	.12	.50 *	.33	.64 **	--							
8. Non-Mutual Tie (Z)	5.00	(2.51)	-.33	-.13	.12	.17	.03	.50 *	.76 ***	--						
9. Density (Z)	0.05	(0.02)	-.05	.12	.12	.50 *	.33	.64 **	1.00 ***	.76 ***	--					
10. Reciprocity (Z)	0.29	(0.30)	.15	.38	-.11	.18	.13	-.08	.11	-.40	.11	--				
11. Diameter (Z)	3.30	(1.22)	-.15	-.44	.22	-.11	.21	.26	.52 *	.67 **	.52 *	-.31	--			
12. Degree-In (Z)	0.12	(0.05)	-.35	.30	-.12	.44	.26	.68 ***	.78 ***	.70 ***	.78 ***	-.03	.26	--		
13. Degree-Out (Z)	0.13	(0.05)	-.12	.39	.10	.47 *	.60 **	.56 **	.65 *	.31	.65 **	.30	.13	.69 ***	--	
14. Betwness (Z)	0.05	(0.04)	-.12	-.11	.02	.23	.13	.54 **	.84 ***	.80 ***	.84 ***	-.05	.68 ***	.72 ***	.46 *	--
15. Mean Distance (Z)	1.82	(0.48)	.00	-.53 *	.18	-.08	.06	.29	.54 *	.68 ***	.54 *	-.30	.95 ***	.26	.07	.74 ***

Note. $N = 20$ fourth-grade students.

Multiple Regressions with Stepwise Predictor Entry

Recall that I used stepwise predictor entry in separate regression models for each item, for each outcome. Logistic regression was used for modeling item performance (1 = correct, 0 = otherwise) and linear regression was used for modeling game engagement (counts of items attempted, post-item, log-transformed). Model results in Table 6 shown that none of the student characteristics were predictive of item performance on any of the items. However, there were a few noteworthy network indices that were related to whether or not students got the item correct on their first attempt, but as expected, those relations were item-specific. Degree-in (ties going into nodes) was the only unique predictor of the first item (Item 511) and although not significant, there were trends for it to negatively predict the fourth item (Item 1062). As well, the number of items practiced prior to the second item (Item 701) was positively predictive of better item performance on that item.

Table 6

Multiple Logistic Regression with Stepwise Predictor Entry Model Results for Item Performance

Parameter	Item 511		Item 701		Item 961		Item 1062	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Conditional Mean (Intcpt)	0.73	(0.52)	**-1.46**	**(0.52)**	**1.70**	**(0.52)**	**-1.22**	**(0.52)**
Items Pre-Item (Log)			**1.07**	**(0.53)**				
Mean RT (Log)								
Mean Ability (Z)								
Item RT (Log)								
Tie Count (Z)								
Non-Mutual Tie (Z)							-1.13	(0.85)
Density (Z)							1.85	(0.96)
Reciprocity (Z)								
Diameter (Z)								
Degree-In (Z)	**1.32**	**(0.53)**					-1.41	(0.88)
Degree-Out (Z)								
Betweenness (Z)								
Mean Distance (Z)								

Note. $N = 20$ fourth-grade students. All values in logits. Boldface indicates significance at the .05 level, 2-tailed.

Last, although none of the indicators were uniquely predictive of Item 961 performance, there were notable trends for three network statistics to predict Item 1062: non-mutual tie (one-directional tie between two nodes) and degree-in centrality were negatively whereas density was positively predictive. In other words, although greater connections among unique actions indicated better item performance, actions that were not reciprocal and actions that were repeatedly used correlated with worse performance.

These findings can be interpreted as, since Item 511 was the first-attempted item by the student players in game, repeated use or exploration of actions might benefit getting familiar with the item as well as the game platform. Since Item 701 was difficult and was attempted in a relatively early phase of the game (recall that it was encountered approximately 3 items after the first), pre-item practice would help increase the chance of getting a correct response. Even though non-mutual tie, density, and degree-in were not predictive of a correct response for Item 1062, the trends suggested that when encountered another difficult item in a relatively later phase of the game (recall that it was encountered approximately 11 items after the first), building more connections among actions as well as building mutually connections among actions would help getting a correct response. However, repeated use of available actions may somehow decrease the probability of getting a correct answer. One possible explanation is that even though even though repeated use of actions may be useful at the beginning stage of the game after a certain number of items have been attempted, repeated use of actions may indicate a lack of understanding or getting stuck, thus negatively impacting item performance.

Turning attention to models of post-item game engagement, I found that student characteristics appeared more important for this outcome (Table 7). For each of the item, student mean ability (measured as percent correct on student first attempts across all items) was

positively predictive of continued game engagement (measured as the count of post-item). Other patterns of associations were item-specific depending on item characteristics.

Table 7

Multiple Linear Regression with Stepwise Predictor Entry Model Results for Game Engagement

Parameter	Item 511		Item 701		Item 961		Item 1062	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Conditional Mean (Intcpt)	**4.31**	**(0.13)**	**4.26**	**(0.15)**	**4.18**	**(0.18)**	**4.05**	**(0.16)**
Items Pre-Item (Log)								
Mean RT (Log)								
Mean Ability (Z)	**0.51**	**(0.14)**	**0.49**	**(0.15)**	**0.50**	**(0.18)**	**0.56**	**(0.16)**
Item RT (Log)					-0.27	(0.18)		
Tie Count (Z)								
Non-Mutual Tie (Z)			**1.03**	**(0.41)**				
Density (Z)								
Reciprocity (Z)								
Diameter (Z)								
Degree-In (Z)	**-0.60**	**(0.17)**			-0.28	(0.18)		
Degree-Out (Z)	0.26	(0.17)						
Betweenness (Z)			-0.68	(0.41)			**0.59**	**(0.24)**
Mean Distance (Z)							**-1.07**	**(0.24)**

Note. $N = 20$ fourth-grade students. All values in counts. Boldface indicates significance at the .05 level, 2-tailed.

For Item 511, degree-in was negatively predictive of continued game engagement while there was a positive trend for degree-out (ties going out of nodes). For Item 701, a markedly more difficult item that was encountered approximately three items after the first, non-mutual tie was positively associated with continued game engagement, with betweenness centrality (the degree to which nodes stand between each other) showed a trend for being positively associated with continued game engagement. For Item 961, a relatively easier item that was encountered after the first two items that was encountered approximately three items after the second, response time on item (measured as total time spent in responding to the item) and degree-in showed trends of negatively associated with continued game engagement. For item 1062, which is another relatively difficult item that was encountered after the first three items that was encountered

approximately three items after the third, betweenness centrality was positively associated but mean distance centrality (the average shortest path between distant nodes) was negatively associated with continued game engagement.

For continued game engagement, first of all, student with higher ability in solving the game items would prefer continuing to attempt future items and getting engaged with the game. Again, since item 511 was the starting item in the game attempted by student players, repeated use of actions would decrease continued engagement in game. For the second item 701, as it was a difficult item attempted at an early phase of the game, more non-mutual ties suggest that fewer back-and-forth movements between actions would help students stay in the game. Both findings indicated that clear subgoals in responding to these two items may motivate students to remain engaged in the game. Furthermore, since the third item 961 was an easy item, more repeated use of actions and more time spent on the item may discourage student's engagement in the game. Finally, since the last item 1062 was another difficult item, the ability to implement effective solutions (i.e., utilizing some actions as a bridge to other while reducing ties between actions) would facilitate continued engagement.

CHAPTER 3.

Aim 2: Using Inferential Network Analyses with Item Action Sequence Data

Beyond descriptive network properties, inferential network approaches can be further utilized to describe patterns of ties within a network. The inferential analysis of networks is a set of theories, models, and applications rather than a simple analytical method (Carolan, 2014; Harris, 2014). One of the most important families of models specifically designed for relational data is exponential random graph models (ERGMs; Holland & Leinhardt, 1981; Leifeld et al., 2018) that attempts to explain how and why network ties arise by predicting the likelihood of a new tie forming between any pair of nodes (also known as dyads), and it can interpret both network- and dyad-level statistics (Cranmer & Desmarais, 2011). In the context of education, ERGMs has been applied to investigate student interactions at school. For example, Gallagher (2015) applied ERGMs in examining students' willingness to communicate between members of different ethnolinguistic groups, and more recently, Yoshida (2022) applied ERGMs in revealing prosocial behaviors reflected by students' exchange of gratitude messages in online learning.

To date, there are no studies that have used ERGMs for analyzing action sequences. Specifically, in Zhu et al.'s (2016) network representation of action sequence data from NAEP items, the extension of applying ERGMs as one of the future directions was discussed. Therefore, the present aim extends the network analysis in Aim 1 from descriptive to inferential network analyses through the application of different approaches under the family of ERGMs to the action networks generated from action sequences of *RiddleBooks* items. Furthermore, the present study focused on combining four types of ERGMs with different levels of student characteristics to explore the patterns of "action tendencies" and to examine the efficiency of these approaches.

Exponential Random Graph Models (ERGMs)

ERGMs can be used to analyze both directed and non-directed, binary and weighted networks, and also to predict the likelihood of a mutual (or *reciprocal*) tie in directed networks for any new tie formation. ERGM benefits from the ease of interpreting results since its parameter interpretation is quite similar as that of a logistic regression (the difference is that ERGMs take the presence or absence of ties as the outcome). For binary networks, parameters are estimated in log-odds units (or *logits*), while for weighted networks, parameters are estimated in counts (Eidsaa & Almaas, 2013; Pilny & Atouba, 2017; Scott, 2015). Furthermore, ERGMs can be estimated using both frequentist estimation approach and Bootstrapped estimation approach (BTERGMs; Camino & Friel, 2011; 2014).

While ERGMs model a single network observed at a single time point, when networks are observed at multiple time points (or repeatedly, as in multiple items), ERGMs can be extended to imply temporal dependencies in two ways. One extension is temporal ERGMs (TERGMs; Robins & Pattison, 2001) that can be estimated using *tergm* or *btergm* packages in *R*. TERGM distinguishes dependencies as within- or cross-time and test the effect by taking into account the network at a prior time point (Leifeld et al., 2018; Schaefer & Marcum, 2017). Temporal ERGM by Bootstrapped (BTERGMs; Krivitsky & Handcock, 2014) shares similarity with TERGM but uses a different estimation method in *R*. It is considered as repeated measures analysis of outcomes for multiple trials (i.e., not growth modeling). The other extension is separated temporal ERGMs (STERGM; Krivitsky & Handcock, 2014) that can be estimated using *tergm* package in *R*. STERGMs manage to separate new tie formation and tie persistence/dissolution processes in network evolution over time (i.e., two models rather than one), and focuses on how different factors might drive each process. It assumes that during a

given discrete time step, the process by which the ties form does not interact with the process by which they dissolve (i.e., both processes are separated from one and another based on the state of the network at the beginning of the time step).

Research Questions

The present study applied and compared four types of ERGMs: 1) ERGMs at binary, individual level, 2) ERGMs at binary, individual level using Bayesian estimation (BERGMs), 3) weighted/valued ERGMs that analyze a valued or count network at aggregated level, and 4) bootstrap methods for temporal ERGMs (BTERGMs) that analyze a grand list of network at aggregated level. Each approach incorporated covariates of student covariates at different levels (i.e., response time at student/item/student-item hybrid levels) and further explored their utility in describing action network structure and predicting action tendencies. I tabulated action sequence and time response data from *RiddleBooks* (the same data as used in Aim 1) in modeling action sequences as networks. Specifically, individual binary, directed networks were created for each student with a starting action (i.e., starting item) and an ending action (i.e., validating model) to for ERGMs and BERGMs, which were then aggregated into a weighted action network (i.e., a count network) for weighted ERGMs. In addition, a joint list of aggregated student networks for each item (i.e., a grand network) was created for BTERGMs. The research questions are:

1. Which network analysis approaches better represent and estimate the action sequence data, especially when student characteristics are added?
2. How can these model results help us predict student performance and engagement using student's repeated and total attempts on items across the entire game?

Method

Data Sources

As the purpose of the present study was to expand on Aim 1 by using inferential models to model item action sequence data to further describe and derive network properties, the same analytic sample and items as Aim 1 were used for this aim, which included $N = 20$ common fourth grade student players attempted four sequential items from *RiddleBooks* (i.e., Items 511(Start), 701, 961, and 1062). Item details can be found in Figure 3. Once again, I limited the action data to first attempts to obtain "naive" item action patterns that were not confounded by that particular item's practice effects.

Measures

Action Sequences and Networks

Similar to Aim 1, each student player's raw time-stamped action sequence of all action patterns possible was mapped into action-by-action directed adjacency matrices for each item, and then transformed to directed, binary network for each individual, which was for individual ERGMs and BERGMs. I further created valued and joint networks for weighted ERGMs and BTERGMs respectively.

Student Characteristics

Student-level predictors. For each item, I created three student-level predictors: 1) total time spent on the item, 2) time spent on each of the individual actions (within student).

Aggregate action-level (item) predictor. For each item, I created an action-level predictor: time spent on each of the actions (across students).

Modeling Approaches

To estimate (measure) the linkages among the action patterns, I utilized four network analysis approaches using the family of exponential random graph models (ERGMs). First, I

used directed, binary ERGM to measure action linkages within each student using the frequentist estimation approach (ERGM; Cranmer & Desmarais, 2011; Handcock et al., 2018) and the Bootstrapped estimation approach (BTERGM; Camino & Friel, 2011; 2014). Both predict the likelihood of a linkage (also known as "tie" or "edge") forming between any pair of actions, and for both models, only one predictor was possible to include in analysis: the student's time spent on each action (within student time on action). I then averaged across models to obtain aggregate results for comparison with subsequent models. The ERGM probability mass function (PMF) is defined as, the probability of observing a tie conditional on the network properties (*x)*, is an exponential function of product of the vector of *k* model parameters to be estimated (denoted θ). A vector of change statistics associated with the parameters and network properties modeled, divided by a normalizing constant, kappa (typically denoted "c") to ensure the model is able to be estimated. The model is estimated in logits form using a maximum likelihood estimation algorithm.

$$Pr_{\theta;\eta,g}(Y = y|x) = \frac{\exp\left(\eta(\theta) * g(y;x)\right)}{\kappa_{\eta,g}(\theta;x)} \quad (1)$$

Next, I conducted a weighted or "valued" ERGM in which student matrices were aggregated into a count matrix for each item; this model predicts the expected count between any pair of actions, and could only include the aggregate time spent across students on each action as a predictor of action linkage/tie (Hunter et al., 2008; Krivitsky, 2014).

Last, I conducted a Temporal Exponential Random Graph Model by Bootstrapped Pseudolikelihood (BTERGM), which allowed us to incorporate both student- and item-level predictors by modeling the probability of a linkage among actions for the last student, taking into account all previous students' action patterns (Krivitsky & Handcock, 2014; Leifeld et al., 2017). This model in particular was of interest as it is much like a repeated measures analysis of

outcomes for multiple trials (i.e., not growth modeling); in this case, trials are students' individual action matrices nested within a larger item-level action matrix.

For all four approaches, all response time predictors of were log-transformed and then standardized into z-scores to facilitate interpretation of results and comparison between modeling methods.

Results

Aggregated Student Network Results

As a reminder, separate ERGMs and BERGMs for each student on each item were conducted to predict the likelihood of a tie forming between any pair of actions, controlling for the relative time spent by the student on a particular action (action level). (The only difference between these two analyses was in the estimation algorithm). The parameter estimate results were then averaged across students (see Table 8, all results in logits).

Table 8

Panel A: Aggregated binary ERGM Results

Effect	**Item 511**		**Item 701**		**Item 961**		**Item 1062**	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	-3.36	(0.64)	-3.81	(0.66)	-3.66	(0.70)	-4.09	(0.80)
Student-Level Time on Action (Log)	0.82	(0.32)	0.79	(0.33)	0.53	(0.40)	1.02	(0.39)

Note. Values are in logits, averaged across $N = 20$ elementary student players' binary directed networks.

Panel B: Aggregated binary BERGM Results

Effect	**Item 511**		**Item 701**		**Item 961**		**Item 1062**	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	-3.61	(0.70)	-4.06	(0.74)	-4.10	(0.56)	-4.38	(0.86)
Student-Level Time on Action (Log)	0.89	(0.34)	0.83	(0.36)	0.52	(0.42)	1.10	(0.42)

Note. Values are in logits, averaged across $N = 20$ elementary student players' binary directed networks.

Translating the logits into probabilities, the mean estimated adjusted probability from the ERGM for Item 511 was 3% for a non-mutual action linkage (tie), and 2% for Items 701, 961 and 1062. Moreover, for each 1 *SD* increase in time a student spent on a given action, there was a

predicted increase of 7% for Item 511 and was a predicted increase of 3%-4% for Items 701, 961, and 1062 in the probability that another action would be linked with that action. The BERGM results were nearly the same, but overall had a slight decrease of 1% in estimating probabilities. Results indicated that as the first Item attempted by student players in the game, Item 511 had the highest predicted probability of non-mutual ties forming between actions. In addition, the more time spent on an action step, the more likely it is that more actions will be utilized in responding to the item.

Item-Level Network Results

The rationale for taking the mean of the parameter estimates above (i.e., aggregating student-level results) was that analyzing just the count data of student linkages at the item level would preclude being able to incorporate students' individual action-level predictors. Nevertheless, analyzing the network data as counts, with number of students who select a particular action sequence as the data, may be useful if student-level data is not of interest. For this analysis, the only predictor available for inclusion is the time spent on each action, aggregated, log-transformed, and z-scored across students.

Results using weighted ERGMs for all four items are shown in Table 9; these results are in logs rather than logits. Translating the logs into counts, for Item 511, I found that the mean estimated number of students who linked any pair of actions was predicted to be 1.08, and the number was much lower for Item 701 (0.82), Item 961 (0.71), and Item 1062 (0.49). Further, for each 1 *SD* increase in time spent on a given action, there was a predicted increase of 1.73 more students forming a link between that action and another action for Items 511, and a predicted increase of 1.20-1.23 for Items 701 and 1062, but only a predicted increase of 0.92 for Item 961.

Although the estimated counts were small, as the first item in the sequence, Item 511 had the highest estimated numbers of action linkages by students.

Table 9

Weighted/Valued ERGM Results

Effect	**Item 511**		**Item 701**		**Item 961**		**Item 1062**	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	0.08	(0.10)	**-0.20**	**(0.09)**	**-0.34**	**(0.11)**	**-0.71**	**(0.13)**
Item-Level Time on Action (Log)	**0.47**	**(0.06)**	**0.38**	**(0.06)**	**0.26**	**(0.08)**	**0.92**	**(0.09)**

Note. Values are in logs; data is counts of $N = 20$ elementary student players in a single weighted directed network. Boldface indicates significant at .05 level.

Student-Item Hybrid Network Results

The most promising results were from the BTERGM model, which is able to incorporate all student-level networks analogous to a repeated measures "trials" type of design that estimates the mean network properties and relationships. For this model, I tested each response time predictor individually as well as combined. Results, shown in Table 10, revealed that time on an action within students was consistently positively predictive of a tie forming in the network for all items, both uniquely in Model 1 and jointly with other predictors in Model 4. Specifically, for each 1 *SD* increase in time a student spent on a given action, there was a predicted increase of 12% for Item 511, a predicted increase of 7%-8% for Items 701 and 961, and a predicted increase of 4% for Item 1062. It is also interesting to note that student total time spent on item was consistently predictive of a tie forming in the network for all items, both uniquely in Model 3 and jointly with other predictors in Model 4. The prediction was positive for all items in Model 3, however, it was only positively associated with Item 1062 while negatively associated with other items in Model 4: for each 1 *SD* increase in total time a student spent on an item, there was a predicted increase of 3% for a tie forming within the network for Item 1062, and a predicted decrease of 1%-3% for the other items. Moreover, although total time spent on action across

students was not consistently predictive across items, it was positively predictive of tie forming within a network for Items 701 and 1062. Recall that, as discussed in Aim 1, since these two items were the most difficult items of the four, the more time spent on a given action across students may indicate a consistent thinking process while figuring out the item, thus more tie formations.

Table 10

BTERGM Results

Effect	**Item 511**							
	Model 1		Model 2		Model 3		Model 4	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	**-3.08**	**(0.17)**	**-2.64**	**(0.11)**	**-2.85**	**(0.17)**	**-3.15**	**(0.16)**
Student-Level Time on Action (Log)	**0.73**	**(0.05)**	--	--	--	--	**1.16**	**(0.13)**
Item-Level Time on Action (Log)	--	--	-0.01	(0.06)	--	--	-0.01	(0.07)
Student-Item Level: Student Total Time on Item (Log)	--	--	--	--	**0.52**	**(0.09)**	**-0.51**	**(0.14)**
Effect	**Item 701**							
	Model 1		Model 2		Model 3		Model 4	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	**-3.43**	**(0.20)**	**-3.09**	**(0.14)**	**-3.15**	**(0.18)**	**-3.62**	**(0.17)**
Student-Level Time on Action (Log)	**0.69**	**(0.05)**	--	--	--	--	**1.07**	**(0.10)**
Item-Level Time on Action (Log)	--	--	**0.21**	**(0.07)**	--	--	**0.24**	**(0.09)**
Student-Item Level: Student Total Time on Item (Log)	--	--	--	--	**0.41**	**(0.04)**	**-0.46**	**(0.09)**
Effect	**Item 961**							
	Model 1		Model 2		Model 3		Model 4	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	**-3.46**	**(0.12)**	**-3.24**	**(0.11)**	**-3.30**	**(0.11)**	**-3.59**	**(0.12)**
Student-Level Time on Action (Log)	**0.51**	**(0.05)**	--	--	--	--	**1.10**	**(0.14)**
Item-Level Time on Action (Log)	--	--	-0.02	(0.08)	--	--	-0.02	(0.08)
Student-Item Level: Student Total Time on Item (Log)	--	--	--	--	**0.27**	**(0.08)**	**-0.67**	**(0.15)**
Effect	**Item 1062**							
	Model 1		Model 2		Model 3		Model 4	
	Coeff	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*	*Coeff*	*(SE)*
Tie (Linkage)	**-3.74**	**(0.14)**	**-3.08**	**(0.11)**	**-3.69**	**(0.15)**	**-3.85**	**(0.16)**
Student-Level Time on Action (Log)	**0.93**	**(0.05)**	--	--	--	--	**0.75**	**(0.07)**
Item-Level Time on Action (Log)	--	--	**0.14**	**(0.06)**	--	--	**0.15**	**(0.07)**
Student-Item Level: Student Total Time on Item (Log)	--	--	--	--	**1.00**	**(0.10)**	**0.32**	**(0.10)**

Note. Values are in logits; data is from $N = 20$ elementary student players in a single grand network list. Boldface indicates significant at .05 level.

CHAPTER 4.

Aim 3: Using Hidden Markov models (HMMs) with Item Action Sequence Data

Yet another novel approach to analyzing process data for psychometric purposes includes Hidden Markov models (HMMs), which have been widely used in sequence analyses (Rabiner, 1989). An HMM consists of two strings of information, including an *observed sequence* and an underlying *non-observed sequence* (i.e., hidden or latent state sequence) derived from the observed sequence (Eddy, 2004; Rabiner & Juang, 1986). HMMs had a long tradition in use to analyze automatic speech recognition (e.g., Gales, 2008; Rabiner, 1989). In the context of educational assessment, HMM is one of the most widely used sequence analysis approaches that models the relationships or dependences among student actions in responding to online tasks. For example, LaMar (2018) applied HMMs in modeling students' configuration actions and the decision to enter subtasks in a cell biology game; Arieli-Attali, Ou, and Simmering (2019) applied HMMs in modeling test taker's choice sequences of item difficulty in a self-adapted test; more recently, Xiao et al. (2021) applied HMMs in modeling examinees' action sequences from a computerized assessment for measuring adult's information processing skills.

Hidden Markov Models (HMMs)

In an HMM, a set of possible latent states can be extracted given an observed sequence as a first-order Markov chain, indicating that what state to transition to next depends only on the current state. Specifically, each observed action in an observed sequence is assumed to be generated by a hidden state according to the response probability distribution of that state. According to Ghahramani (2001): setting the sequence length as *T,* the joint likelihood of observed actions $Y_{1:T} = (Y_1, \dots, Y_T)$ and latent states $S_{1:T} = (S_1, \dots, S_T)$, can be written as

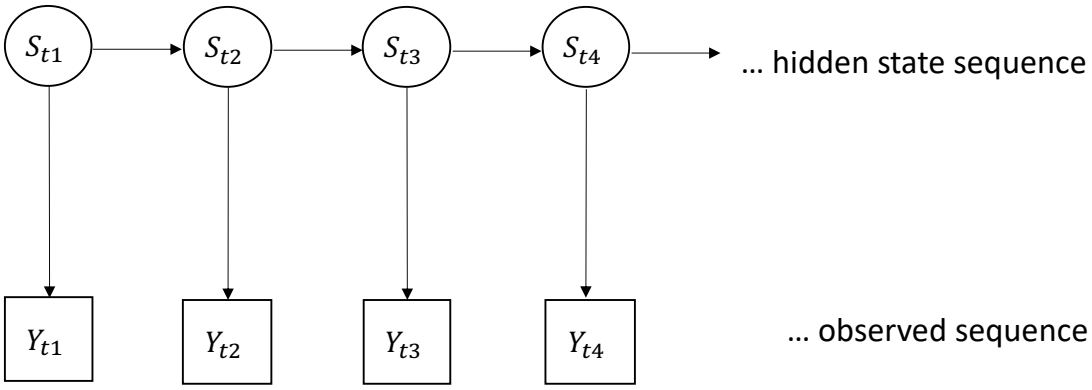
$$(Y_{1:T}, S_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T-1} P(S_t|S_{t-1})P(Y_t|S_t) \quad (2)$$

where S_t belongs to a set of k latent states, $P(S_1)$ represents the probability of the initial state, $P(Y_t|S_t)$ represents the conditional probability of the action Y_t given the latent state S_t (also called "emission probability" or "response probability"), and $(S_t|S_{t-1})$ represents the transition probability between states, independent to the time point t. Both initial state and transition probabilities may depend on the estimated response probabilities. The models were estimated through the expectation–maximization (EM) algorithm given the observed action sequences and the number of possible states that were pre-defined. Model comparisons were implemented through AIC and BIC model fit indices, where a lower value of AIC or BIC indicates a better model fit. The model fit information can help to determine the best optimal number of latent states that can best represent the data.

Overall, there are three sets of parameters of interest: 1) initial state probability indicating the starting point of the item, 2) response/emission probabilities indicating latent states generated from observed sequences, and 3) transition probabilities indicating transition paths among the latent states. Figure 4 shows the relationship between the latent/hidden state sequence and the observed action sequence: the circles represent the latent states at different time points while the squares represent the observed action(s); the arrows pointing from circles to squares represent the conditional probabilities that the latent state evokes the observation, and the arrows among circles of latent states represent how latent states transit to others. Thus, by estimating latent states and latent state transitions, we are able to make inferences about the problem-solving patterns underlying the observed action sequences.

Figure 5

Path Diagram of Relationship between Hidden and Observed Sequences in HMMs



While the first two aims were focused on extracting information from action networks in predicting student outcomes, the present aim shifts focus to examining behavioral patterns in responding to items by taking account into student performance level. For this aim, I focus on the action sequence data from the starting item (Item 511) in *RiddleBooks* teacher's version which had the most student information available in examining problem-solving behaviors for students. The research questions are:

1. What are certain patterns of behavior for students with different levels of game performance?
2. What is the potential for providing feedback to task developers from the model results?

Method

Data Sources

In this study, I utilized the action sequence data from the starting item in *RiddleBooks* teacher's version (i.e., Item 511), which had the most user information available (with $N = 639$ students). After data cleaning (removal of actions less than 2 in the first attempts), $N = 556$ student players' action sequences were used in the final study. Since I was also interested in student players' action patterns across response groups, student players were divided into two subgroups: correct = 1 and incorrect = 0, with $n = 341$ and $n = 215$, respectively. Once again, a

correct response represents whether a student player validated the bar model successfully in the first attempt.

Action Sequences

The action sequences for all players had an average action count of 17.15 (*SD* = 9.89), followed a right-skewed distribution (shown in Figure 5). Since I was also interested in student players' action patterns across response groups, action sequences were compared across the two subgroups (i.e., correct = 1 vs. incorrect = 0). Specifically, the correct response group had an average action count of 17.72 (*SD* = 9.56) while the incorrect response group had an average action count of 16.25 (*SD* = 10.35). Figure 6 shows the boxplots of the action sequence lengths for the two groups. Overall, student players who answered Item 511 correctly performed slightly more actions compared with those who answered Item 511 incorrectly. Additionally, the total number of unique actions used by correct and incorrect groups were 18 and 13, respectively, which can be interpreted as, student players who answered Item 511 correctly tried out more unique actions compared with those who answered Item 511 incorrectly.

Figure 6

Item 511 Action Sequence Lengths

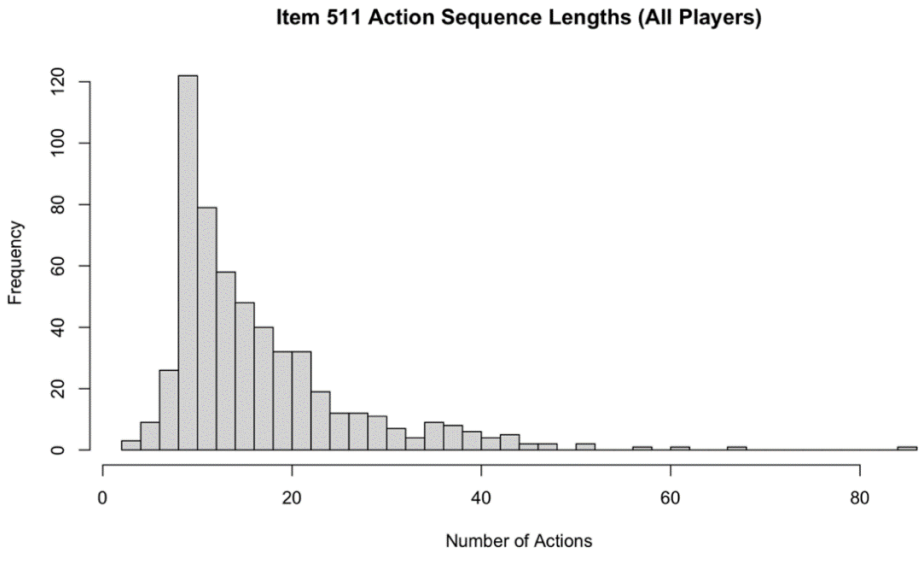
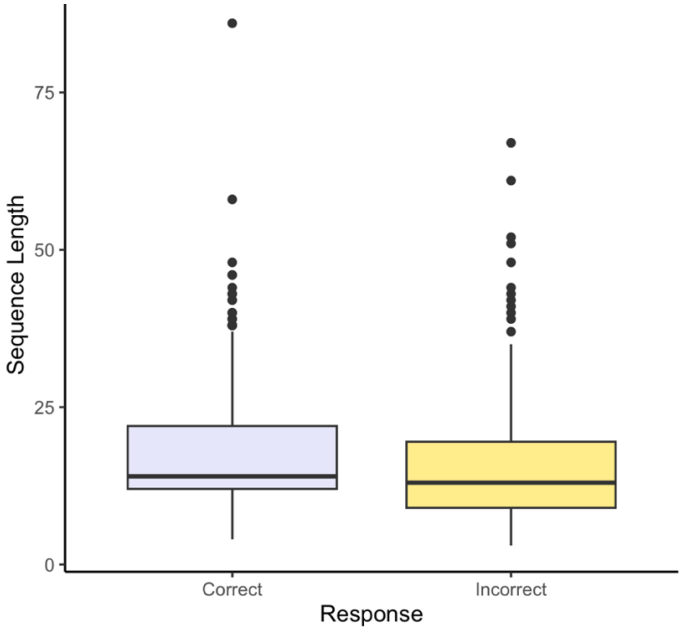


Figure 7

Item 511 Action Sequence Lengths by Response Groups



Modeling Approaches

Three sets of HMMs were run separately for each response group (all responses, correct responses, and incorrect responses) with pre-defined numbers of states. After that, the model of best fit with a certain number of states was selected for each response group based on model fit indices. Model results were thus extracted from the selected models for each response group separately. Specifically, for each response group, the labels for each latent state were summarized based on the estimated actions and the corresponding response probabilities within each state, and the latent state transitions were visualized using R *igraph* package (Csardi & Nepusz, 2006).

Results

Model Fit

Table 11 shows the model indices of AIC and BIC for all pre-defined models with various state numbers for three response groups (i.e., all responses, correct responses, incorrect

responses). The optimal number of latent states were decided at a 6-state model for each response group. The latent states and transitions can therefore be extracted from the selected models, separately for each group. In general, the model results provide us: 1) the predicted probability for initial state, 2) the predicted actions and the corresponding response probabilities in the latent states, and 3) the predicted transition probabilities between the latent states. The initial state tells us the starting point of the sequences, and labels for each latent state was summarized based on the certain action types and predicted response probabilities. Specifically, for each latent state, the labels are summarized by more "dominant" actions (i.e., actions with relatively larger estimated response probabilities in the state). After summarizing the state labels from the predicted response probabilities, we can further investigate the transition probabilities among the labeled action states, thus indicating the problem-solving patterns underlying the latent states and their transitions.

Table 11

HMM Model Fit Indices

Model Comparison	All Responses $N = 556$		Correct Responses $N = 341$		Incorrect Responses $N = 215$	
	AIC	BIC	AIC	BIC	AIC	BIC
1-state model	28076.22	28197.99	18304.59	18418.60	9676.04	9749.94
2-state model	27635.04	27900.07	18055.55	18303.70	8569.54	8735.83
3-state model	24646.15	25068.77	16250.91	16646.61	8515.07	8786.06
4-state model	24251.56	24846.10	15822.08	16378.74	8398.23	8786.23
5-state model	22994.69	23775.46	15464.85	16195.89	8304.79	8822.13
6-state model	**22537.89**	**23519.23**	**14620.16**	**15539.00**	**7849.68**	**8508.67**
7-state model	22764.10	23960.32	14902.47	16022.51	8134.52	8947.48

Note. Lower AIC and BIC indicate better fit; Indices in bold are the model of selections.

Latent States

The initial action state was estimated as the action of “Pickup Information” from the item so that student players could begin to build the bar model. Even though a 6-state was decided as

the optimal number of states in representing the input action sequence data, the details of the actions and predicted response probabilities were different. Table 12 shows a heatmap of the distributions of actions and their corresponding response probabilities estimated from the models. To emphasize, the results were extracted from a separate HMM for each response group. Specifically, the light to dark purple color in the heatmap reflects how high the predicted response probability is for a given action grouped in the state (i.e., the darker the color, the higher the probability). Action IDs represent actions with different purpose in responding to the item (again, see Appendix A for details). There's no order among the state id numbers, and there can be one or more actions estimated for each latent state, depending on the specific estimated response probabilities. If a state had only one action with high response probability, the it was labeled with the conceptual meaning of this action; if a state had multiple actions with high response probability, then it was labeled after the combination of the conceptual meaning of these actions (as shown in Appendix C).

Overall, the All Response and the Correct Response groups had more single-dominant actions within states while the Incorrect Response group had more multiple-dominant actions with states. To map the actions and response probabilities with conceptual meanings, the actions with the largest response probability within each action state are summarized in Table 13. It is interesting to notice that the Incorrect Response group had two states with "Pickup Information" as the largest response probability within each state, while there's no state with "Validate Model" as the largest response probability. This indicates that student players who answered the item incorrectly had more actions in picking up information pieces (i.e., phrases) from the item content, but may lose interest or had trouble in validating the model.

Table 12

Summary of Action States from Response Probabilities

New Action	Raw Action	All Responses						Correct Responses						Incorrect Responses					
		St1	St2	St3	St4	St5	St6	St1	St2	St3	St4	St5	St6	St1	St2	St3	St4	St5	St6
1	1&2			0.01	0.06	0.87	0.23	0.02	0.86					0.26		0.95	0.01	0.02	0.78
2	17			0.35	0.08	0.08	0.01		0.08	0.31		0.01		0.46				0.39	0.01
3	19			0.37	0.06	0.05	0.01		0.05	0.37	0.01			0.13		0.02		0.35	
4	20			0.19	0.77					0.21		0.99		0.12				0.14	
5	3	0.78						0.73			0.05				0.01	0.01	0.68	0.06	
6	4		0.83	0.02	0.02		0.02			0.01			0.98	0.02	0.87				
7	10				0.01		0.01	0.01		0.00			0.02	0.01					
8	9		0.01		0.01			0.01											
9	12																		
10	5												0.01						
11	11																		
12	25	0.03	0.02					0.03			0.02				0.02		0.03	0.01	
13	30			0.02						0.01	0.01			0.01	0.01			0.03	
14	24	0.10	0.11	0.03				0.20		0.07				0.01	0.08		0.09	0.01	0.01
15	18	0.01								0.01							0.01		
16	23			0.01						0.01									
17	28			0.01						0.01									
18	31	0.08	0.03				0.71				0.91				0.01	0.01	0.19		0.21

Note. Response probabilities were estimated from the model of selections (6-state models) for each group.

Table 13

Actions with the Highest Response Probability for Each State by Group

State	All Responses	Correct Responses	Incorrect Responses
St1	Pickup Expression	Pickup Expression	Add Label to Model
St2	Drop Expression	Pickup Information	Drop Expression
St3	Add Bar to Model	Add Bar to Model	Pickup Information
St4	Add Label to Model	Validate Model	Pickup Expression
St5	Pickup Information	Add Label to Model	Add Bar to Model
St6	Validate Model	Drop Expression	Pickup Information

State Transitions

The predicted transition probabilities between latent states were visualized as path diagrams to illustrate the state transitions using *R* igraph package (Csardi & Nepusz, 2006; see Figure 6), which allow us to identify the underlying problem-solving behaviors across response groups (only transition probabilities larger than 0.10 are shown in graphs; original transition matrix tables are shown in Appendix B). Again, state labels were re-summarized by the actions and response probabilities estimated for each state (as shown in Appendix C).

The All Response group showed a very high transition probability of the Pickup Information state staying in itself (0.86) and only a very low transition probability to other states, suggesting that there was a lot of repeated use of grabbing information from the item context. However, results would be re-interpreted if comparing the two response subgroups. By looking at the transition patterns of the two response groups, only the Incorrect Response group had a high transition probability (0.56) of staying within the Pickup Information state, suggesting repeated use of item information happened much more for student players who answered the item incorrectly.

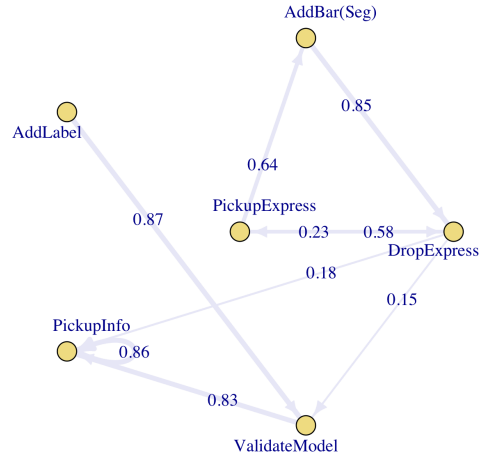
In addition, even though the transition probability from the Pickup Information state to the Validate Model state was very low (0.02) and almost identical to the Correct Response

group, the direct transition from the Pickup Information state to the Pickup Information/Validation model state was very large for the Incorrect Response group (0.97). This indicates that student players who answered the item incorrectly may have quickly skipped or lost interest (i.e., tended to end after playing with the item information phrases). Moreover, the Correct Response group was shown to have clear subgoals in the problem-solving process, such as transitioning directly from the Pickup Information state to the Add Label state (i.e., utilizing a certain phrase to attach a label to represent the math relationship of the bar model). However, this finding was not shown in the All Response group or the Incorrect Response group. Therefore, the findings were more informative by dividing students by item responses.

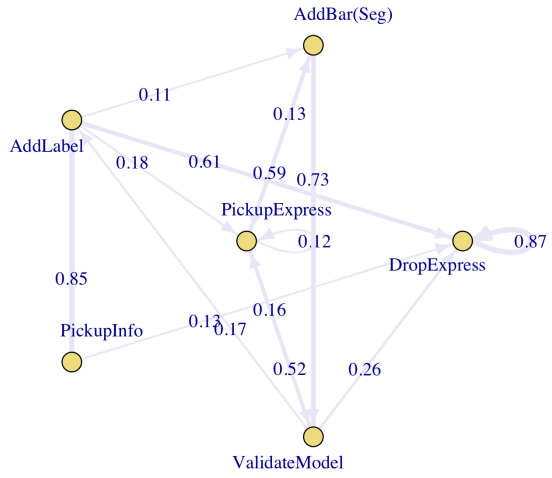
Figure 8

Path Diagram of Results of HMM for Item 511

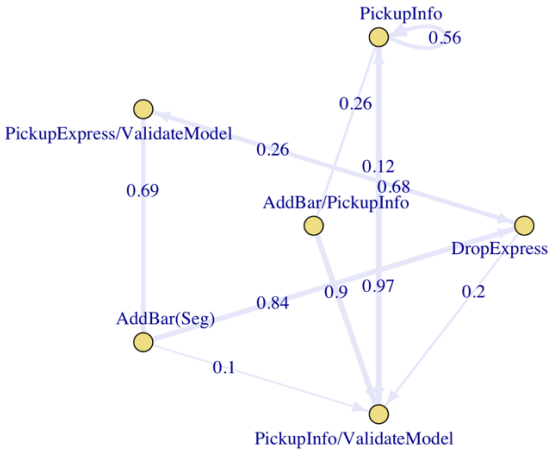
Panel A: All Responses



Panel B: Correct Responses



Panel C: Incorrect Responses



CHAPTER 5.

Aim 4: Using Latent Class Analysis (LCA) with Item Action Sequence Data

For Aim 4 I again employed data from the first item of *RiddleBooks* to evaluate the extent to which item action sequence data could produce latent classes (clusters) of students, which in turn might potentially also be able to predict item performance and post-item game engagement, above and beyond response time. To this end, a Latent Class Analysis (LCA) was performed using *Mplus* (Muthén & Muthén, 1998/2017). Specifically, LCA has been widely used by educational and psychological researchers to identify subgroups (i.e., latent classes or hidden groups) that are similar to each other from observed responses, for example, to examine acculturation patterns of ethnic minorities (Jang et al., 2017). LCA is a cross-sectional model with categorical outcomes and can be thought of as a "person-centered" approach (Nylund-Gibson & Choi, 2018). To be clear, the previous Aim 3 also utilizes a form of latent class analysis, but that aim investigates clusters of action sequence transition patterns; in Aim 4, I simplify the analysis to extract latent classes of students using a set of action-sequence combination indicators without imposing a transition structure.

The LCA model assumes that the response pattern on a set of indicators (p observed binary variables, u_1, u_2, u_3, …, u_p) measured by N individuals is reflective of an underlying unordered categorical latent variable, c, that has K latent classes (e.g., Nylund-Gibson & Choi, 2018). In an unconditional model (i.e., without covariates), the parameters of interest in LCA results include the number of classes, the relative size (proportion) of the classes, and the indicator thresholds for each class. Importantly, LCA is focused on how indicator patterns are a result of a heterogeneous (mixed) population; however, it does not assume that individuals only belong to one class – model results can be used to assign each individual a predicted probability

of belonging to each of the classes. Moreover, an LCA can incorporate both covariates and outcomes in a single model. The present study utilized LCA in modeling action sequence data along with the position of actions within each sequences (i.e., "steps/moves") as combinations. The research questions are:

1. What are the action sequence patterns that give rise to different classes? Are different classes associated with different levels of game performance and engagement?
2. What is the potential for providing feedback to task developers from the model results?

Method

Data Sources

Like Aim 3, I used the action sequence data from the first attempt of $N = 500$ fourth graders to complete the fourth-grade level Item 511 (the first item administered in the game). For this Aim 4, each student's individual action data were arranged into "steps" (or "moves"), after beginning the item. Students varied in the total number of steps they took, ranging from 1 to 28 ($M = 21.59$, $SD = 13.72$, *Median* = 18). Within each step, I coded the top seven most frequent types of actions across all students and time logs as a nominal scale, ranging from 1 = Action 1&2 (Pickup & Drop Information Phrases), 2 = Action 17 (Add Bar), 3 = Action 19 (Add Bar Segment), 4 = Action 20 (Add Label), 5 = Action 3 (Pickup Expression), 6 = Action 4 (Drop Expression), 7 = Action 24 (Remove Bar From Model), and used 8 to represent all of the less frequent actions. Next, for each step individually (across students), I collapsed categories that occurred for less than 5% of students into an "other" category = 9 to improve model estimation, especially given that each category within a step requires additional parameters. Last but not

least, because students completed at different numbers of steps, I coded the "completion" action two ways for analyses: as either an additional category = 10, or simply as "missing".

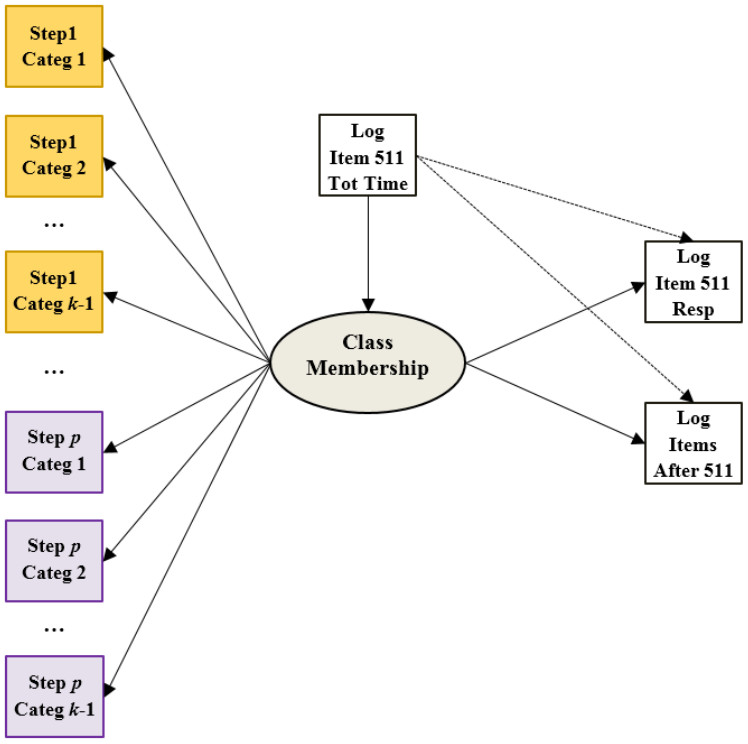
The predictor, **total time spent** answering Item 511, was heavily right-skewed (*Skew* = 9.40) and leptokurtotic (*Kurtosis* = 145.30), so consistent with the previous analyses, I transformed it as the natural logarithm + 1 (**logTotTime**). Similarly, the **engagement outcome**, the number of items attempted after Item 511, was heavily skewed (*Skew* = 5.32) and leptokurtotic (*Kurtosis* = 46.36), so again we transformed it by taking the natural logarithm + 1 (**logItemCt**). Our other outcome, item performance, was binary where 1 = student answered the item correctly on the first attempt and 0 = otherwise.

Data Analyses

The first stage in the data analysis was to determine the number of classes for the action sequence data, and in so doing, determine whether to model the completion action as another category or as a missing category. I specified 1-, 2-, and 3-class mixture models for each type of completion action handling, and used fit indices and a likelihood ratio test for model selection. The second stage in the analysis was to specify a latent class structural regression model in which I used class membership, as well as total time on the item, to predict student both outcomes – whether or not the students got the item correct, and the number of items they engaged in going forward (see Figure 7 for general path diagram). All models were estimated using *Mplus 8* (Muthén & Muthén, 1998/2017). For this stage, I compared two models: a model in which total time on the item had the same relationship with the outcomes across classes, and a model in which total time's predictive relationship with the outcomes was free to vary within classes. Across all models, each step was treated as a nominal scale of different types of actions.

Figure 9

LCA Model for Step-Action Category Data for Item 511



Results

Number of classes. Fit indices comparing 1-, 2-, and 3-class models for the step-action nominal-scale data are provided in Table 14. Irrespective of how completion action was coded, the 2-class model fit significantly better than the 1-class model (LRT *p*s < .001), whereas the 3-class model did not fit significantly differently from the 2-class model (LRT *p*s > .700). Therefore, I extracted step-action threshold results from the 2-class model as provided in Table 15, with "competition" coded as its own action. Both original Action IDs and data Action IDs re-coded for the convenience of analyses are shown in the columns, corresponding to each step-actions numbers (the conceptual meanings of actions can be found in the action coding book in Appendix A). Specifically, Action 1 (Pickup & Drop Information Phrases) was predicted to be used in steps 1-12 in Class 1 while it was predicted to be used in steps 14-19 in Class 2, which

indicates that Class 1 utilized this action more and earlier than Class 2 when answering the item. Additionally, Action 7 (Remove Bar from Model) was not predicted to occur in any of the steps. This is because, if students believed they made the correct decisions to add bars, then deleting bars was not necessary to build the bar models. Similarly, all the less frequent actions in Action 8 were not predicted to occur at any step. Other actions were predicted to occur at some steps in both classes, however there was no consistent pattern across classes.

Table 14

Model Results Comparing Number of Classes Estimated by Type of Completion Action Handling

Completion Action Coded…	Index	1-Class Model	2-Class Model	3-Class Model	1-Class vs. 2-Class	2-Class vs. 3-Class
… as Category[a]	Num Parms	66	133	200	67	67
	LogLik	-9089	-7023	-6367	2066	656
	Deviance (-2LL)	18179	14046	12734	**-4133**	-1312
	AIC	18311	14312	13134	-3999	-1178
	BIC	18589	14872	13977	-3717	-895
	aBIC	18379	14450	13342	-3929	-1108
	Mean Sensitivity	--	1.000	.999	--	-0.001
	Mean Specificity	--	1.000	1.000	--	0.000
	Entropy	--	1.000	0.998	--	-0.002
… as Missing[b]	Num Parms	48	97	146	49	49
	LogLik	-5130	-4466	-4136	664	330
	Deviance (-2LL)	10259	8932	8272	**-1327**	-660
	AIC	10355	9126	8564	-1229	-562
	BIC	10558	9535	9179	-1023	-356
	aBIC	10405	9227	8716	-1178	-511
	Mean Sensitivity	--	.976	.964	--	-0.012
	Mean Specificity	--	.975	.982	--	0.007
	Entropy	--	.904	.927	--	0.023

Note. N = 500 students' data from 19 steps for Item 511. Num Parms = number of parameters estimated; LogLik = the value of the log-likelihood at final iteration; Deviance = -2*LogLik (distributed approximately chi-square and used in the likelihood ratio test; LRT); AIC = Aikake's information criterion; BIC = Bayesian information criterion; aBIC = sample-size adjusted BIC; Mean Sensitivity = mean of average posterior probability of cases assigned to each class being from the same class (ranging from 0 to 1, with higher values being better); Mean Sensitivity = 1 – mean of the average posterior probability of cases assigned to each class being from different class(es) (ranging from 0 to 1, with higher values being better); Entropy = overall measure posterior probability of cases assigned to correct classes (ranging from 0 to 1, with higher values being better). [a] The 2-class model was significantly better than the 1-class model (LRT $p < .001$), with $n = 294$ in class 1 and $n = 206$ in class 2. The 3-class model did not fit significantly better than the 2-class model (LRT $p = .772$), with counts of $n = 175$, $n = 167$, and $n = 158$ in each class.[b] The 2-class model was significantly better than the 1-class model (LRT $p < .001$), with $n = 300$ in class 1 and $n = 200$ in class 2. The 3-class model did not fit significantly better than the 2-class model (LRT $p = .760$), with counts of $n = 207$, $n = 138$, and $n = 155$ in each class.

Table 15

LCA Step-Action Threshold Results for the 2-Class Model with Completion as its own Action

Step: Action Number	Original Action ID	Data Action ID	*Count*	*(%)*	Class 1				Class 2			
					Coeff	*(SE)*	*Z*	*p*	*Coeff*	*(SE)*	*Z*	*p*
S1:1	1&2	1	495	(.99)	4.28	(0.503)	8.51	<.001	5.32	(1.002)	5.31	<.001
S1:2	other	9	5	(.01)								
S2:1	1&2	1	166	(.33)	3.35	(0.719)	4.66	<.001	--	--	--	--
S2:2	17	2	324	(.65)	4.74	(0.710)	6.67	<.001	14.87	(0.140)	106.26	<.001
S2:3	other	9	8	(.02)	1.25	(0.802)	1.56	.118	10.31	(1.005)	10.26	<.001
S2:4	complete	10	2	(.00)								
S3:1	1&2	1	488	(.98)	4.56	(0.580)	7.86	<.001	--	--	--	--
S3:2	other	9	9	(.02)	0.29	(0.764)	0.38	.706	11.31	(0.453)	24.97	<.001
S3:3	complete	10	3	(.01)								
S4:1	1&2	1	408	(.82)	3.70	(0.413)	8.96	<.001	--	--	--	--
S4:2	17	2	60	(.12)	1.67	(0.445)	3.76	<.001	13.23	(0.204)	64.71	<.001
S4:3	other	9	26	(.05)	0.77	(0.494)	1.57	.117	12.46	(0.288)	43.25	<.001
S4:4	complete	10	6	(.01)								
S5:1	1&2	1	219	(.44)	1.97	(0.308)	6.39	<.001	--	--	--	--
S5:2	17	2	25	(.05)	0.15	(0.393)	0.39	.695	12.51	(0.314)	39.87	<.001
S5:3	19	3	194	(.39)	2.55	(0.300)	8.52	<.001	13.80	(0.180)	76.52	<.001
S5:4	3	5	33	(.07)	0.51	(0.365)	1.40	.162	12.68	(0.291)	43.62	<.001
S5:5	other	9	17	(.03)	-0.41	(0.456)	-0.89	.374	12.31	(0.344)	35.73	<.001
S5:6	complete	10	12	(.02)								
S6:1	1&2	1	371	(.74)	1.80	(0.180)	10.01	<.001	--	--	--	--
S6:2	19	3	24	(.05)	-0.88	(0.307)	-2.85	.004	12.17	(0.343)	35.47	<.001
S6:3	4	6	27	(.05)	-0.88	(0.307)	-2.85	.004	12.45	(0.300)	41.55	<.001
S6:4	other	9	42	(.08)	-1.28	(0.357)	-3.58	<.001	13.44	(0.194)	69.12	<.001
S6:5	complete	10	36	(.07)								
S7:1	1&2	1	284	(.57)	1.42	(0.181)	7.85	<.001	--	--	--	--
S7:2	19	3	62	(.12)	0.17	(0.220)	0.77	.443	12.99	(0.258)	50.30	<.001
S7:3	3	5	39	(.08)	-0.64	(0.276)	-2.32	.020	13.10	(0.246)	53.26	<.001
S7:4	4	6	29	(.06)	-0.93	(0.305)	-3.05	.002	12.80	(0.282)	45.44	<.001
S7:5	other	9	48	(.10)	-0.69	(0.281)	-2.47	.014	13.52	(0.206)	65.71	<.001
S7:6	complete	10	38	(.08)								
S8:1	1&2	1	213	(.43)	0.34	(0.164)	2.09	.037	--	--	--	--
S8:2	17	2	32	(.06)	-1.02	(0.243)	-4.21	<.001	12.39	(0.345)	35.87	<.001
S8:3	19	3	31	(.06)	-1.39	(0.280)	-4.96	<.001	12.90	(0.273)	47.15	<.001
S8:4	20	4	92	(.18)	0.31	(0.165)	1.86	.062	11.80	(0.456)	25.86	<.001
S8:5	3	5	36	(.07)	-1.86	(0.340)	-5.46	<.001	13.45	(0.216)	62.29	<.001
S8:6	4	6	24	(.05)	-2.77	(0.515)	-5.38	<.001	13.18	(0.241)	54.68	<.001
S8:7	other	9	8	(.02)	--	--	--	--	12.27	(0.365)	33.62	<.001
S8:8	complete	10	64	(.13)								
S9:1	1&2	1	203	(.41)	-0.66	(0.135)	-4.85	<.001	--	--	--	--
S9:2	19	3	24	(.05)	-2.44	(0.279)	-8.74	<.001	12.52	(0.329)	38.02	<.001
S9:3	3	5	27	(.05)	-5.08	(1.003)	-5.06	<.001	13.47	(0.216)	62.27	<.001
S9:4	4	6	36	(.07)	-2.19	(0.249)	-8.79	<.001	13.10	(0.253)	51.84	<.001
S9:5	other	9	50	(.10)	-2.19	(0.249)	-8.79	<.001	13.68	(0.199)	68.75	<.001
S9:6	complete	10	160	(.32)								

Table 15, Cont'd.

Step: Action Number	Original Action ID	Data Action ID	*Count*	*(%)*	Class 1				Class 2			
					Coeff	*(SE)*	*Z*	*p*	*Coeff*	*(SE)*	*Z*	*p*
S10:1	1&2	1	174	(.35)	-1.41	(0.170)	-8.29	<.001	--	--	--	--
S10:2	20	4	59	(.12)	-1.16	(0.154)	-7.53	<.001	11.51	(0.508)	22.68	<.001
S10:3	4	6	40	(.08)	-2.87	(0.325)	-8.82	<.001	13.53	(0.202)	66.83	<.001
S10:4	other	9	51	(.10)	-2.87	(0.325)	-8.82	<.001	13.84	(0.179)	77.33	<.001
S10:5	complete	10	176	(.35)								
S11:1	1&2	1	153	(.31)	-1.86	(0.177)	-10.53	<.001	--	--	--	--
S11:2	3	5	25	(.05)	--	--	--	--	13.47	(0.221)	61.07	<.001
S11:3	other	9	84	(.17)	-2.53	(0.238)	-10.60	<.001	14.42	(0.155)	93.08	<.001
S11:4	complete	10	238	(.48)								
S12:1	1&2	1	127	(.25)	-4.44	(0.581)	-7.64	<.001	--	--	--	--
S12:2	20	4	35	(.07)	-2.10	(0.190)	-11.03	<.001	11.57	(0.508)	22.77	<.001
S12:3	3	5	27	(.05)	--	--	--	--	13.48	(0.212)	63.45	<.001
S12:4	4	6	23	(.05)	-4.44	(0.581)	-7.64	<.001	13.18	(0.241)	54.68	<.001
S12:5	other	9	35	(.07)	-4.15	(0.504)	-8.23	<.001	13.61	(0.201)	67.80	<.001
S12:6	complete	10	253	(.51)								
S13:1	1&2	1	120	(.24)	--	--	--	--	--	--	--	--
S13:2	3	5	25	(.05)	--	--	--	--	13.43	(0.220)	61.09	<.001
S13:3	4	6	23	(.05)	--	--	--	--	13.35	(0.228)	58.64	<.001
S13:4	other	9	38	(.08)	--	--	--	--	13.85	(0.186)	74.41	<.001
S13:5	complete	10	294	(.59)								
S14:1	1&2	1	97	(.19)	--	--	--	--	2.50	(0.368)	6.78	<.001
S14:2	20	4	26	(.05)	--	--	--	--	1.18	(0.404)	2.92	.004
S14:3	4	6	30	(.06)	--	--	--	--	1.32	(0.398)	3.32	.001
S14:4	other	9	45	(.09)	--	--	--	--	1.73	(0.384)	4.50	<.001
S14:5	complete	10	302	(.60)								
S15:1	1&2	1	92	(.18)	--	--	--	--	0.65	(0.178)	3.65	<.001
S15:2	other	9	66	(.13)	--	--	--	--	0.32	(0.190)	1.68	.093
S15:3	complete	10	342	(.68)								
S16:1	1&2	1	62	(.12)	--	--	--	--	0.07	(0.183)	0.37	.715
S16:2	other	9	86	(.17)	--	--	--	--	0.39	(0.170)	2.32	.020
S16:3	complete	10	352	(.70)								
S17:1	1&2	1	58	(.12)	--	--	--	--	-0.42	(0.169)	-2.47	.014
S17:2	other	9	60	(.12)	--	--	--	--	-0.38	(0.167)	-2.29	.022
S17:3	complete	10	382	(.76)								
S18:1	1&2	1	42	(.08)	--	--	--	--	-0.85	(0.184)	-4.59	<.001
S18:2	4	6	24	(.05)	--	--	--	--	-1.41	(0.228)	-6.18	<.001
S18:3	other	9	42	(.08)	--	--	--	--	-0.85	(0.184)	-4.59	<.001
S18:4	complete	10	392	(.78)								
S19:1	1&2	1	37	(.07)	--	--	--	--	-1.18	(0.188)	-6.26	<.001
S19:2	other	9	49	(.10)	--	--	--	--	-0.90	(0.170)	-5.28	<.001
S19:3	complete	10	414	(.83)								

Note. All values in logits compared to (last) reference category within step and class (significant positive values indicate a probability greater than .50 compared to reference category).

Prediction of student outcomes. Two models were specified for the structural regression stage of analysis, with both including (the log of) the total time spent on the item as a predictor of class membership and student item performance (at the first attempt of the item, which was the first item in the game). Specifically, one model constrained the predictor – outcome relationship to be the same across classes, and the second allowed the relationship to differ within each class. A comparison of the regression-related portion of the models is given in Table 16. Although the fit indices did not substantially differ between the two models, the results show that parameter estimates do differ between the two classes. Specifically, in the constrained model, total time spent on item was negatively predictive of item correct in both classes (*Coeff* = -0.43, $p < .05$).

Table 16

LCA Regression Parameter Results for Constrained and Unconstrained Models

Parameter		Constrained Model				Unconstrained Model			
		Coeff	*(SE)*	*Z*	*p*	*Coeff*	*(SE)*	*Z*	*p*
Overall *Intercepts:*									
Logit(Class 1)		6.84	(0.89)	7.65	<.001	6.84	(0.89)	7.65	<.001
Slopes:									
Log(Item Count)	→ Logit(Class 1)	-1.81	(0.25)	-7.20	<.001	-1.81	(0.25)	-7.20	<.001
Class 1 *Intercepts:*									
Log(Item Count)		2.99	(0.32)	9.26	<.001	2.75	(0.37)	7.37	<.001
Logit(Item Correct = 1)		-1.87	(0.60)	-3.12	.002	-2.18	(0.75)	-2.89	.004
Slopes:									
Log(Item Total Time)	→ Log(Item Count)	**-0.16**	**(0.09)**	**-1.68**	**.094**	**-0.08**	**(0.11)**	**-0.77**	**.440**
Log(Item Total Time)	→ Logit(Item Correct = 1)	**-0.43**	**(0.17)**	**-2.49**	**.013**	**-0.53**	**(0.22)**	**-2.38**	**.017**
Class 2 *Intercepts:*									
Log(Item Count)		3.07	(0.36)	8.44	<.001	3.58	(0.67)	5.37	<.001
Logit(Item Correct = 1)		-2.16	(0.69)	-3.16	.002	-1.54	(1.12)	-1.38	.169
Slopes:									
Log(Item Total Time)	→ Log(Item Count)	**-0.16**	**(0.09)**	**-1.68**	**.094**	**-0.29**	**(0.17)**	**-1.66**	**.096**
Log(Item Total Time)	→ Logit(Item Correct = 1)	**-0.43**	**(0.17)**	**-2.49**	**.013**	**-0.27**	**(0.29)**	**-0.94**	**.347**
Fit Index		*Value*				*Value*			
Num Parms		141				143			
LogLik		-8070				-8069			
Deviance (-2LL)		16139				16138			
AIC		16421				16424			
BIC		17016				17026			
aBIC		16568				16573			

However, in the unconstrained model, total time spent on item was only negatively predictive of item response for Class 1 (*Coeff* = -0.53, $p < .05$) but not Class 2. Converting logits to probabilities, in the unconstrained model, for every 1 *SD* increase in total time spent on item, the probability of getting a correct response decreased by 6%.

Results from both thresholds of step-actions and regression predictions indicated that total time spent on item significantly differentiated the two classes of students. Recall that Action 1 (Pickup & Drop Information Phrases) is a necessary and critical step in start building the bar model as well as in the process of building the bar model; therefore, Class 1 would be considered to include students who performed typical problem-solving patterns with more actions of grabbing information occurring in the first set of steps (1-12). Furthermore, more time spent on item indicated possible hesitations or lack of goals and therefore a lower probability of a correct answer. However, in contrast to Class 1, Class 2 would be considered to include students who engage in exploratory or innovative (or possibly confused) problem-solving patterns, where actions of grabbing information occurred in the very last set of steps (14-19), and in this class, more time spent on the item had no predictive effect on the probability of a correct answer.

CHAPTER 6.

Discussion

This dissertation demonstrated four novel analytic approaches for exploring patterns in action sequence process data from a game-based assessment that can be useful for psychometric research – specifically in terms of trying to better understand why examinees respond to items correctly or incorrectly, as well as what motivates students continue to attempt future items when given the choice (e.g., engagement). Specifically, the first two aims focused on treating types of actions as part of an unordered (but directed) "network" and then quantifying those network properties in different ways to predict student item performance and post-item engagement (aim 1), as well as to predict action-to-action linkage formation within items (aim 2). In contrast, the last two aims incorporate the order of the sequences to explore the usefulness of latent class modeling approaches in predicting within-item action transitions (aim 3) and student outcomes (aim 4). The first two aims extended previous exploratory network descriptive research by Zhu et al. (2016) by using a network perspective to analyze action sequence data. Specifically, in aim 1, I used a 2-step descriptive-prediction analysis, and in aim 2, I used exponential random graph models (ERGMs) that were able to incorporate different levels of item and student predictor characteristics into the modeling process directly. Finally, the last two aims extended recent work by He et al. (2021) in demonstrating ways in which to employ latent class analysis (LCA) modeling for investigating action sequence patterns (rather than action-to-action noise like the first two aims), including estimating *clusters of within-item action transitions* (aim 3) as well as estimating *clusters of students* based on similarities among action sequence patterns (aim 4). In aim 3, the focus is on a specific item characteristic (e.g., starting order of the item) whereas in aim 4, the focus was on multiple student outcomes.

These aims and their corresponding approaches are in fact interrelated. For the first two aims, even the same set of network data were used in analyses, aim 1 emphasized the network's descriptive statistics in explaining and predicting student outcomes, while aim 2 extended the descriptive level network statistics to inferential analytic models and demonstrated the possibility of modeling action sequences as networks to predict possible change of network structures that indicates students' response patterns, especially when student characteristics information was included. As action networks provide us with information about action linkages at the individual and observed level (i.e., an action network was directly transformed from an observed action sequence for each student), aim 3 afforded us the opportunity to examine the latent action sequence underlying the observed action sequence, taking into account all students' observed action sequences together. In this way, HMMs investigated latent classes (states) of actions and transition patterns given the input observed action sequences, which provided us with understanding of students' response patterns at the latent level. Aim 3 results lead naturally to aim 4, which estimated latent classes (clusters) of students given the input of observed action-steps combinations (in a simplified manner, without transition patterns). Below I summarize findings for each aim individually.

Aim 1. Descriptive network statistics and student characteristics showed more impact on student game engagement compared with item performance. However, the impact was item-specific. It was interesting to find that for the starting item (Item 511, moderate difficulty), more repeated actions used when answering the starting item would benefit students in getting a correct answer on this item, but not continued engagement. For an earlier-presented difficult item (Item 701, second item), students who had fewer back-and-forth actions would prefer to continue playing the game. Meanwhile, pre-item practice may benefit students in getting a correct answer.

For a later-presented difficult item (Item 1062, forth item), students who implemented more effective solutions would prefer to continue playing the game. The only consistent finding across items is that student ability was positively predictive of continued engagement: students with higher math modeling ability would prefer to engage more in the game.

Aim 2. In terms of model comparison, binary ERGMs and BERGMs produced similar model results in predicting the probability of s action linkages, and time spent on action within student was the only predictor to be incorporated. Weighted/valued ERGMs provided a different perspective by interpreting action linkages by counts. However, since this approach used one single weighted network of counts on action linkages across students, aggregated time spent on action across students was the only predictor to be incorporated. Finally, as model used for longitudinal analyses of social network data, BTERGMs was able to incorporate all student/item-level predictors into one model in predicting the probability of student's action linkages, which hold great potential for evaluating student-level linkages among actions within an item.

In terms of model results, weighted ERGMs showed that for all items, the more time spent on a given action across students, the more students from action linkages. However, the strength of effect varied by item: the starting item (Item 511, moderate difficult) had the largest effect, while the last item (Item 961, easy) had the smallest effect. In addition, when time spent on a given action across students used as a joint predictor in BTERGMs, it only had a positive predictive effect on students' action linkages for the difficult items (Item 701, second item; Item 961, forth item). Even though the difference was tiny, the earlier-presented item (Item 701) had a larger effect than the later-presented item (Item 961). Moreover, for all items, more time spent on an action, more action linkages form in the future, especially for the starting item (Item 511). Finally, as more time spent on item, it was less likely that action linkages would be formed for

the first three items (Items 511, 701, and 961) but more likely that action linkages would be formed for the last item (Item 1062, difficult). Therefore, similar as Aim 1, students' problem-solving patterns may vary depending on the location and difficulty of a certain item.

Aim 3. In the results explored in Aims 1 and 2, I found that the first attempts on the very first item in the game (Item 511) contained more rich information compared to the other items. Therefore, Aims 3 and 4 focused on this starting item to demonstrate potential latent clustering approaches in further depth. Specifically, Aim 3 focused on the latent clustering of actions by students. Through HMMs on observed action sequences in responding to an item, latent states and transitions by item response groups (correct vs. incorrect) were able to be compared and visualized. These findings revealed that the optimal number of latent states was decided at 6-state for all groups (i.e., all responses, correct responses, incorrect responses). However, the details of actions and corresponding probabilities were different within each action state/clustering for different groups, with the incorrect response group had the most chaotic combinations of observed actions without a solid ending point, suggesting more random actions performed by the incorrect response group and possible skip without validating the bar model. Furthermore, by examining the latent transitions, the correct response group showed the signs of clear subgoals while the incorrect response group showed more repeated actions. It is also worth noting that dividing students by response groups was more efficient than a single model.

Aim 4. The use of LCA for the initial game item's action sequences shows promise in classifying students and predicting student outcomes, including both item performance and future game engagement (as measured by number of items attempted after the first). Even though the latter two aims both fall within the scope of latent class analysis, a distinction should be stated: whereas the former aim was to reveal student behavioral patterns at the latent stage by

examining the latent sates and transition patterns, this aim extracted latent classes of students by estimating patterns of step-action combinations and was simplified in the absence of transition patterns. Results showed that the optimal number of students classes was decided as two regardless of how completion action was coded (missing/other). Furthermore, incorporating response time has succeeded in differentiating student groups.

Limitations and Future Directions

The major limitation of this study is that real data, rather than simulated data, was employed; as such, we cannot know how these models would perform relative to a true population model with varied conditions. In a similar vein, since this dissertation is focused on online education game data rather than traditional assessment data, the sample is quite unique. In particular, as the game was designed for elementary school students, the items were relatively easy and did not require many action steps in answering the items. Therefore, how these approaches would perform with varied sample sizes, number of action sequences, and sparseness levels is yet unknown. Future directions of research should include evaluating these analytic approaches using simulated data with different conditions, such as simulating action sequences with various numbers of unique steps and actions that may or may not need to be collapsed, as well as different levels of action-to-action interactivity (i.e., by designing data networks with different densities and degrees of reciprocity).

Again, in terms of the sample characteristics, the items were administered in an unsupervised order, rather than every student being administered every item; the action networks did not incorporate the idea of action ordering but only the action connections. Therefore, future research may include the order of action linkages as predictor in obtaining more detailed results.

Moreover, the fact that the 2-class model fit the data best in Aim 4 indicates that perhaps those classes could be used as predictors in the BTERGMs in Aim 2.

Another limitation includes the use of response data from users' first attempts: this does not allow for understanding in the learning process for each successive attempt, which is another avenue that could be studied (e.g., use of the "hint" function). It would be interesting to include the usage of the game functions in the future, as well as other attempts after their first attempts to reveal more details about students' problem-solving strategies. In addition, even though practice effects were considered (as measured as the count of items attempted before the focal item), this predictor did not provide information about the *total* amount of practice, as students could have multiple attempts on each item before progressing to another item.

Last but not least, this particular game environment also had another unique feature: it was an *unsupervised* learning tool, which is problematic in terms of how the engagement outcome was defined in these analyses. Specifically, I used the total number of items attempted after the focal item; however, this implicitly assumes that all students had the same number of opportunities to access more items, yet this is likely not the case. Again, future research with other types of assessment environments, and with greater information available around the gaming decision tree process, is necessary for future work on how these novel analytic approaches can be useful for understanding respondents' thinking as well as assessment improvements.

Practical Implications

On a practical level, the results of this study can shed some light on students' math problem-solving strategies, especially when student-level characteristics were incorporated. For example, results from Aim 1 (descriptive network approach) can provide feedback on how

actions were connected to each other. For example, repeated use of certain actions may indicate possible struggles if network degree negatively predicts item performance, problem-solving strategy would be efficient if diameter negatively predicts item performance. Results from Aim 2 (inferential network approach) would provide feedbacks through the likelihood prediction of action tie formation give the amount of response time. For example, more actions would happen in the future if response time were positively predicted of tie formation, which is also supported by common knowledge. Therefore, when there's an opposite prediction, it may be interesting to double check the item content to avoid any problematic issues. In addition, both Aims 1&2 found that student or network characteristics were associated with item properties, such as the positions of items and varying difficulty level of items. From the perspective of test development, it would be helpful to figure out an efficient item delivery model that would help student players stay interested in the game, especially if they are not getting the correct answers and positive feedback. However, since order of actions in network models were not incorporated as a predictor, the results from the network analyses may not provide sufficient guidance on where the linkages occurred and to what extent that the order matters, while latent class analyses in Aims 3 and 4 included such order information, which have the potential to provide more reliable information for test developers. Results from Aim 3 provided insights on how student problem-solving behaviors differ by item response group, whereas results from Aim 4 distinguished students by response time. These findings may be helpful to detect problematic items if students' responses to the same item vary significantly.

Conclusion

Process data – the ancillary information generated during an assessment – has the potential to provide researchers with rich information, especially within-item, time-stamped

action sequence patterns, and may thus provide assessment developers with a better understanding of examinees' thinking processes (or problematic item instructions, stems, or response choices!). In this study, action sequence data from an online math game environment was successfully analyzed with four novel analytic approaches, with the first two focusing on network analyses and the second two focusing on latent class analyses. Through the network approaches, action sequences were successfully quantified as action networks, and descriptive network statistics and inferential network models were then computed to predict student outcomes. Through latent class approaches, clusters of action patterns as well as clusters of students were successfully extracted from the observed action sequences; in the latter analyses, those clusters were significantly predictive of student outcomes. Each approach was shown to have both strengths and weaknesses in attempting to glean insights about the cognitive processes underlying examinee problem-solving patterns as they attempt items with different degrees of difficulty and timing. Such insights have the potential to be extended to other sequential process data analysis and further contribute to test development and test validity.

References

Arieli-Attali, M., Ou, L., & Simmering, V. R. (2019). Understanding test takers' choices in a self-adapted test: A hidden Markov modeling of process data. *Frontiers in Psychology*, *10*, 83–83. https://doi.org/10.3389/fpsyg.2019.00083

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732. https://doi.org/10.3102/1076998618784700

Bergner, Y., Shu, Z., & von Davier, A. (2014). *Visualization and Confirmatory Clustering of Sequence Data from a Simulation-based Assessment Task*. Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014).

Caimo, A., & Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, *33*(1), 41–55. https://doi.org/10.1016/j.socnet.2010.09.004

Caimo, A., & Friel, N. (2014). Bergm: Bayesian exponential random graph models in R. *Journal of Statistical Software*, *61*(2), 1-25. https://doi.org/10.18637/jss.v061.i02

Carolan, B. V. (2014). *Social network analysis and education: Theory, methods & applications*. SAGE Publications. https://doi.org/10.4135/9781452270104

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(2), 195-212. https://doi.org/10.1007/BF01246098

Center for Game Science. (2015). *Riddlebooks*. A math educational game funded by National Science Foundation (1546510). The University of Washington.

Chen, Y., Zhang, J., Yang, Y., & Lee, Y. S. (2022). Latent space model for process data. *Journal of Educational Measurement*, *59*(4), 517-535. https://doi.org/10.1111/jedm.12337

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486. https://doi.org/10.3389/fpsyg.2019.00486

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education*, *42*(3), 309-328. https://doi.org/10.1080/15391523.2010.10782553

Cranmer, S. J., & Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Political Analysis*, *19*(1), 66-86. https://doi.org/10.1093/pan/mpq037

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695*. https://igraph.org/

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H. C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology, 6*, 348. https://doi.org/10.3389/fpsyg.2015.00348

Debrenti, E. (2015). Visual representations in mathematics teaching: An experiment with Students. *Acta Didactica Napocensia, 8*(1), 19-25.

Dutilh, G., Annis, J., Brown, S.D., et al. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review, 26*, 1051-1069. https://doi.org/10.3758/s13423-017-1417-2

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology, 22*(10), 1315–1316. https://doi.org/10.1038/nbt1004-1315

Eidsaa, M., & Almaas, E. (2013). s-core network decomposition: A generalization of k-core analysis to weighted networks. *Physical Review. E*, *88*(6), 062819–062819. https://doi.org/10.1103/PhysRevE.88.062819

Farahani, F. V., Karwowski, W., & Lighthall, N. R. (2019). Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Frontiers in Neuroscience*, *13*, 585. https://doi.org/10.3389/fnins.2019.00585

Gales, M. (2008). The application of hidden Markov models in speech recognition. In *Foundations and Trends in Signal Processing* (Vol. 1, Issue 3, pp. 195–304). Now Publishers. https://doi.org/10.1561/2000000004

Gallagher, H. C., & Robins, G. (2015). Network statistical models for language learning contexts: Exponential random graph models and willingness to communicate. *Language Learning*, *65*(4), 929–962. https://doi.org/10.1111/lang.12130

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(1), 9–42. https://doi.org/10.1142/S0218001401000836

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1*(1), 103-111. https://doi.org/10.1016/j.dadm.2014.11.003

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*, 108-119. https://doi.org/10.1016/j.intell.2011.02.001

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement, 15,* 128-132. https://doi.org/10.1080/15366367.2017.1411651

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Handcock M., Hunter D., Butts C., Goodreau, S., Krivitsky, P., & Morris, M. (2018). *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org). R package version 3.9.4, https://CRAN.R-project.org/package=ergm.

Harris, J. K. (2014). *An Introduction to Exponential Random Graph Modeling* (Vol. 173, Quantitative Applications in the Social Sciences). Sage Publications.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., Borgonovi, F., & Suárez ´Alvarez, J. (2022). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning, 7*, 1-18. https://doi.org/10.1111/jcal.12748

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, *76*(373), 33-50. https://doi.org/10.1080/01621459.1981.10477598

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, *24*(3). https://doi.org/10.18637/jss.v024.i03

Hwang, G. J., Wang, S. Y., & Lai, C. L. (2021). Effects of a social regulation-based online learning framework on students' learning achievements and behaviors in mathematics. *Computers & Education, 160*, 104031. https://doi.org/10.1016/j.compedu.2020.104031

Jiao, H., He, Q., & Veldkamp, B. P. (2021). Process data in educational and psychological measurement. *Frontiers in Psychology, 12*, 793399. https://doi.org/10.3389/fpsyg.2021.793399

Krivitsky, P. N., & Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society*. *Series B*, *Statistical Methodology, 76*(1), 29-46. https://doi.org/10.1111/rssb.12014

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika*, *83*(1), 67–88. https://doi.org/10.1007/s11336-017-9570-0

Lazega, E., & Snijders, T. A. B. (2016). *Multilevel network analysis for the Social Sciences: Theory, methods and applications*. Springer.

Leifeld, P., Cranmer, S. J., & Desmarais, B. A. (2018). Temporal Exponential Random Graph Models with btergm : Estimation and Bootstrap Confidence Intervals. *Journal of Statistical Software, 83*(6), 1–36. https://doi.org/10.18637/jss.v083.i06

Liao, D., He, Q., & Jiao, H. (2020). Using log files to identify sequential patterns in PIAAC problem solving environments by US adults' employment-related variables. *International Review of Education, 54*(5-6), 627-650.

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 261-288. https://doi.org/10.1111/bmsp.12175

Muthén, L. K., & Muthén, B. O. (1998/2017). *Mplus User's Guide (8th Ed.)*. Los Angeles, CA: Muthén & Muthén.

Naldi, L., & Cazzaniga, S. (2020). Research techniques made simple: Latent class analysis. *Journal of Investigative Dermatology*, *140*(9), 1676-1680. https://doi.org/10.1016/j.jid.2020.05.079

Nasrinpour, H. R., Friesen, M. R., D, R., & McLeod. (2016). *An Agent-Based Model of Message Propagation in the Facebook Electronic Social Network*. https://doi.org/10.48550/arxiv.1611.07454

Ng, S. F., & Lee, K. (2009). The model method: Singapore children's tool for representing and solving algebraic word problems. *Journal for Research in Mathematics Education*, *40*(3), 282–313. https://doi.org/10.5951/jresematheduc.40.3.0282

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440–461. https://doi.org/10.1037/tps0000176

Pilny, A., & Atouba, Y. (2017). Modeling Valued Organizational Communication Networks Using Exponential Random Graph Models. *Management Communication Quarterly, 32*(2), 250-264. https://doi.org/10.1177/0893318917737179

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. https://doi.org/10.1109/MASSP.1986.1165342

Reis Costa, D., Bolsinova, M., Tijmstra, J., & Andersson, B. (2021a). Improving the precision of ability estimates using time-on-task variables: Insights from the PISA 2012 computer-based assessment of mathematics. *Frontiers in psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.579128

Robins, G., Pattison, P., & Elliott, P. (2001). Network models for social influence processes. *Psychometrika*, *66*(2), 161-189. https://doi.org/10.1007/BF02294834

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education*, *8*(1), 5. https://doi.org/10.1186/s40536-020-00082-1.

Schaefer, D. R., & Marcum, C. S. (2017). Modeling network dynamics. In *The Oxford Handbook of Social Networks* (pp. 254-287). https://doi.org/10.1093/oxfordhb/9780190251765.013.19

Scott, T. A. (2015). Analyzing policy networks using valued exponential random graph models: Do government-sponsored collaborative groups enhance organizational networks? *Policy Studies Journal*, *44*(2), 215–244. https://doi.org/10.1111/psj.12118

Scrimgeour, M. B., & Huang, H. H. (2022). A Comparison of paper-based and computer-based formats for assessing student achievement. *Mid-Western Educational Researcher*, *34*(1).

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling*, *59*(1), 109.

Thompson, J., Richards, J., Shim, S.-Y., Lohwasser, K., Von Esch, K. S., Chew, C., Sjoberg, B., & Morris, A. (2019). Launching networked PLCs: Footholds into creating and improving knowledge of ambitious and equitable teaching practices in an RPP. *AERA Open*, *5*(3), 233285841987571. https://doi.org/10.1177/2332858419875718

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425-453. https://doi.org/10.1080/00273171.2019.1643699

Valente, T. W., Gallaher, P., & Mouttapa, M. (2004). Using social networks to understand and prevent substance use: A transdisciplinary perspective. *Substance Use & Misuse*, *39*(10–12), 1685–1712. https://doi.org/10.1081/ja-200033210

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287-308. https://doi.org/10.1007/s11336-006-1478-z

Vankúš, P. (2021). Influence of game-based learning in mathematics education on students' affective domain: A systematic review. *Mathematics*, *9*(9), 986. https://doi.org/10.3390/math9090986

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Visser, I., & Speekenbrink, M. (2010). depmixS4: an R-package for hidden Markov models. *Journal of Statistical Software*, *36*(7), 1–21. https://doi.org/10.18637/jss.v036.i07

White, D. R. (2014). Kinship, class, and community. In J. Scott & P. J. Carrington (Eds.), *The SAGE Handbook of Social Network Analysis* (pp. 129-147). SAGE Publications.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878–903. https://doi.org/10.1002/sce.21027

Xiao, Y., He, Q., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Journal of Computer Assisted Learning*, *37*(5), 1232–1247. https://doi.org/10.1111/jcal.12559

Xie, B., Davidson, M., Li, M., & Ko, A. (2019a). An item response theory evaluation of a language-independent CS1 knowledge assessment. *Proceedings of the 50th ACM*

Technical Symposium on Computer Science Education, 699–705. https://doi.org/10.1145/3287324.3287370

Yoshida, M. (2022). Network analysis of gratitude messages in the learning community. *International Journal of Educational Technology in Higher Education*, *19*(1), 47–47. https://doi.org/10.1186/s41239-022-00352-8

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for Educational Assessment. *Journal of Educational Measurement*, *53*(2), 190–211. https://doi.org/10.1111/jedm.12107

Appendices

Appendix A

Table A1

Details of Action Interpretations and Coding Book

Action.id	Content
1	PHRASE_PICKUP_EVENT
2	PHRASE_DROP_EVENT
3	EXPRESSION_PICKUP_EVENT
4	EXPRESSION_DROP_EVENT
17	ADD_NEW_BAR
18	ADD_NEW_BAR_COMPARISON
19	ADD_NEW_BAR_SEGMENT
20	ADD_NEW_HORIZONTAL_LABEL
21	ADD_NEW_VERTICAL_LABEL
22	ADD_NEW_UNIT_BAR
23	REMOVE_BAR_COMPARISON
24	REMOVE_BAR_SEGMENT
25	REMOVE_HORIZONTAL_LABEL
26	REMOVE_VERTICAL_LABEL
28	RESIZE_HORIZONTAL_LABEL
30	SPLIT_BAR_SEGMENT
35	MUTIPLY_BAR
31	VALIDATE_BAR_MODEL

Appendix B

Table B1

Details of Transition Probabilities for HMM (Aim 3)

Transition Probabilities between Actions States (All Responses)

State	ToS1	ToS2	ToS3	ToS4	ToS5	ToS6
FromS1: Pickup Expression	0.03	0.23	0.64	0.03	0.06	0.00
FromS2: Drop Expression	0.58	0.05	0.02	0.02	0.18	0.15
FromS3: Add Bar (Segment)	0.05	0.85	0.02	0.00	0.07	0.02
FromS4: Add Label	0.09	0.03	0.01	0.00	0.00	0.87
FromS5: Pickup Information	0.04	0.01	0.01	0.07	0.86	0.02
FromS6: Validate Model	0.09	0.03	0.02	0.02	0.83	0.01

Note. Results were estimated from 556 student players who attempted Item 511.

Transition Probabilities between Actions States (Players with Correct Responses)

State	ToS1	ToS2	ToS3	ToS4	ToS5	ToS6
FromS1: Pickup Expression	0.12	0.08	0.59	0.01	0.05	0.16
FromS2: Pickup Information	0.04	0.87	0.00	0.01	0.06	0.01
FromS3: Add Bar (Segment)	0.13	0.06	0.04	0.05	0.00	0.73
FromS4: Validate Model	0.18	0.61	0.11	0.00	0.02	0.08
FromS5: Add Label	0.00	0.13	0.00	0.85	0.00	0.02
FromS6: Drop Expression	0.52	0.26	0.03	0.17	0.00	0.03

Note. Results were estimated from 341 student players who answered Item 511 correctly.

Transition Probabilities between Actions States (Players with Incorrect Responses)

State	ToS1	ToS2	ToS3	ToS4	ToS5	ToS6
FromS1: Add Bar/Pick up Information	0.03	0.00	0.00	0.07	0.00	0.90
FromS2: Drop Expression	0.01	0.04	0.04	0.68	0.03	0.20
FromS3: Pickup Information	0.26	0.01	0.56	0.05	0.01	0.12
FromS4: Pickup Expression/Validate Model	0.00	0.26	0.00	0.05	0.69	0.00
FromS5: Add Bar (Segment)	0.00	0.84	0.03	0.01	0.02	0.10
FromS6: Pickup Information/Validate Model	0.03	0.00	0.97	0.00	0.00	0.00

Note. Results were estimated from 215 student players who answered Item 511 incorrectly.

Appendix C

Table C1

Details of State Labels Summarized from Response Probabilities for HMM (Aim 3)

State ID	State Label	Response Probabilities
St1	Pickup Expression	EXPRESSION_PICKUP_EVENT (0.78); REMOVE_BAR_SEGMENT (0.10); VALIDATE_BAR_MODEL (0.08)
St2	Drop Expression	EXPRESSION_DROP_EVENT (0.83); REMOVE_BAR_SEGMENT (0.11)
St3	Add Bar (Segment)	ADD_NEW_BAR_SEGMENT (0.37); ADD_NEW_BAR (0.35); ADD_NEW_HORIZONTAL_LABEL (0.19)
St4	Add Label	ADD_NEW_HORIZONTAL_LABEL (0.77); ADD_NEW_BAR (0.08); PHRASE_PICKUP&DROP_EVENT (0.06); ADD_NEW_BAR_SEGMENT (0.06)
St5	Pickup Information	PHRASE_PICKUP&DROP_EVENT (0.87); ADD_NEW_BAR (0.08)
St6	Validate Model	VALIDATE_BAR_MODEL (0.71); PHRASE_PICKUP&DROP_EVENT (0.23)

Note. Results were estimated from 556 student players who attempted Item 511; Response probabilities larger than 0.05 are shown in table.

Summary of Action States from Response Probabilities (Players with Correct Responses)

State ID	State Label	Response Probabilities
St1	Pickup Expression	EXPRESSION_PICKUP_EVENT (0.73), REMOVE_BAR_SEGMENT (0.20)
St2	Pickup Information	PHRASE_PICKUP&DROP_EVENT (0.86), ADD_NEW_BAR (0.08)
St3	Add Bar (Segment)	ADD_NEW_BAR_SEGMENT (0.37), ADD_NEW_BAR (0.31), ADD_NEW_HORIZONTAL_LABEL (0.21), REMOVE_BAR_SEGMENT (0.08)
St4	Validate Model	VALIDATE_BAR_MODEL (0.91)
St5	Add Label	ADD_NEW_HORIZONTAL_LABEL (0.99)
St6	Drop Expression	EXPRESSION_DROP_EVENT (0.98)

Note. Results were estimated from 341 student players who answered Item 511 correctly; Response probabilities larger than 0.05 are shown in table.

Summary of Action States from Response Probabilities (Players with Incorrect Responses)

State ID	State Label	Response Probabilities
St1	Add Bar/Pick up Information	ADD_NEW_BAR (0.46), PHRASE_PICKUP&DROP_EVENT (0.26), ADD_NEW_BAR_SEGMENT (0.13); ADD_NEW_HORIZONTAL_LABEL (0.12)
St2	Drop Expression	EXPRESSION_DROP_EVENT (0.87), REMOVE_BAR_SEGMENT (0.08)
St3	Pickup Information	PHRASE_PICKUP&DROP_EVENT (0.95)
St4	Pickup Expression/Validate Model	EXPRESSION_PICKUP_EVENT (0.68), VALIDATE_BAR_MODEL (0.19), REMOVE_BAR_SEGMENT (0.09)
St5	Add Bar (Segment)	ADD_NEW_BAR (0.39), ADD_NEW_BAR_SEGMENT (0.35), ADD_NEW_HORIZONTAL_LABEL (0.14), EXPRESSION_PICKUP_EVENT (0.06)
St6	Pickup Information/Validate Model	PHRASE_PICKUP&DROP_EVENT (0.78), VALIDATE_BAR_MODEL (0.21)

Note. Results were estimated from 215 student players who answered Item 511 incorrectly; Response probabilities larger than 0.05 are shown in table.

Appendix D

Aim 1 R Code: Network Statistics & Stepwise Regressions (Item 511 as Example)

```
# Network Statistics for One Individual
gsize.511 <- gsize(ntwk.511) # tie count
non.m.511 <- unname(table(which_mutual(ntwk.511))[1] # non-mutual tie count
density.511 <- edge_density(ntwk.511) # density
reciprocity.511 <- reciprocity(ntwk.511) # reciprocity
diameter.511 <- diameter(ntwk.511, directed=T) # diameter
din.511 <- centr_degree(ntwk.511, mode="in", normalize=T)$centralization # degree-in
dout.511 <- centr_degree(ntwk.511, mode="out", normalize=T)$centralization # degree-out
betw.511 <- centr_betw(ntwk.511, directed=T, normalize=T)$centralization # betweenness
mean_distance.511 <- mean_distance(ntwk.511,directed=T) # mean distance

# Logistic regression (outcome = item correct) ----
fullModel.511 = glm(Item.response ~ ., family = 'binomial', data = stepwise.511.binary,control =
       list(maxit = 2000, epsilon=1)) # all variables
nullModel.511 = glm(Item.response ~ 1, family = 'binomial', data = stepwise.511.binary, control =
       list(maxit = 2000, epsilon=1)) # intercept only
summary(stepAIC(nullModel.511, # start with a model containing no variables
                direction = 'forward', # run forward selection
                scope = list(upper = fullModel.511, # the maximum with all variables
                             lower = nullModel.511), # the minimum with no variables
                trace = 0)) # do not show the step-by-step process of model selection

# Linear regression (outcome = # post-item) ----
fullModel.511.after = lm(LogAfter.Item511 ~ .,  data = stepwise.511.after) # all variables
nullModel.511.after = lm(LogAfter.Item511 ~ 1,  data = stepwise.511.after) # intercept only
summary(stepAIC(nullModel.511.after, # start with a model containing no variables
                direction = 'forward', # run forward selection
                scope = list(upper = fullModel.511.after, # the maximum to consider with all
       variables
                             lower = nullModel.511.after), # the minimum to consider with no
       variables
                trace = 0)) # do not show the step-by-step process of model selection
```

Appendix E

Aim 2 R Code: ERGMs with Covariates (Item 511 as Example)

```
# Binary ERGMs ----
WInet1.fit.511 = ergm(ntwk.511.1 ~ edges + nodecov("within_student"))
WInet2.fit.511 = ergm(ntwk.511.2 ~ edges + nodecov("within_student"))
WInet3.fit.511 = ergm(ntwk.511.3 ~ edges + nodecov("within_student"))
WInet4.fit.511 = ergm(ntwk.511.4 ~ edges + nodecov("within_student"))
WInet5.fit.511 = ergm(ntwk.511.5 ~ edges + nodecov("within_student"))
WInet6.fit.511 = ergm(ntwk.511.6 ~ edges + nodecov("within_student"))
WInet7.fit.511 = ergm(ntwk.511.7 ~ edges + nodecov("within_student"))
WInet8.fit.511 = ergm(ntwk.511.8 ~ edges + nodecov("within_student"))
WInet9.fit.511 = ergm(ntwk.511.9 ~ edges + nodecov("within_student"))
WInet10.fit.511 = ergm(ntwk.511.10 ~ edges + nodecov("within_student"))
WInet11.fit.511 = ergm(ntwk.511.11 ~ edges + nodecov("within_student"))
WInet12.fit.511 = ergm(ntwk.511.12 ~ edges + nodecov("within_student"))
WInet13.fit.511 = ergm(ntwk.511.13 ~ edges + nodecov("within_student"))
WInet14.fit.511 = ergm(ntwk.511.14 ~ edges + nodecov("within_student"))
WInet15.fit.511 = ergm(ntwk.511.15 ~ edges + nodecov("within_student"))
WInet16.fit.511 = ergm(ntwk.511.16 ~ edges + nodecov("within_student"))
WInet17.fit.511 = ergm(ntwk.511.17 ~ edges + nodecov("within_student"))
WInet18.fit.511 = ergm(ntwk.511.18 ~ edges + nodecov("within_student"))
WInet19.fit.511 = ergm(ntwk.511.19 ~ edges + nodecov("within_student"))
WInet20.fit.511 = ergm(ntwk.511.20 ~ edges + nodecov("within_student"))

# Binary BERGMs ----
WIBnet1.fit.511 = bergm(ntwk.511.1 ~ edges + nodecov("within_student"))
WIBnet2.fit.511 = bergm(ntwk.511.2 ~ edges + nodecov("within_student"))
WIBnet3.fit.511 = bergm(ntwk.511.3 ~ edges + nodecov("within_student"))
WIBnet4.fit.511 = bergm(ntwk.511.4 ~ edges + nodecov("within_student"))
WIBnet5.fit.511 = bergm(ntwk.511.5 ~ edges + nodecov("within_student"))
WIBnet6.fit.511 = bergm(ntwk.511.6 ~ edges + nodecov("within_student"))
WIBnet7.fit.511 = bergm(ntwk.511.7 ~ edges + nodecov("within_student"))
WIBnet8.fit.511 = bergm(ntwk.511.8 ~ edges + nodecov("within_student"))
WIBnet9.fit.511 = bergm(ntwk.511.9 ~ edges + nodecov("within_student"))
WIBnet10.fit.511 = bergm(ntwk.511.10 ~ edges + nodecov("within_student"))
WIBnet11.fit.511 = bergm(ntwk.511.11 ~ edges + nodecov("within_student"))
WIBnet12.fit.511 = bergm(ntwk.511.12 ~ edges + nodecov("within_student"))
WIBnet13.fit.511 = bergm(ntwk.511.13 ~ edges + nodecov("within_student"))
WIBnet14.fit.511 = bergm(ntwk.511.14 ~ edges + nodecov("within_student"))
WIBnet15.fit.511 = bergm(ntwk.511.15 ~ edges + nodecov("within_student"))
WIBnet16.fit.511 = bergm(ntwk.511.16 ~ edges + nodecov("within_student"))
WIBnet17.fit.511 = bergm(ntwk.511.17 ~ edges + nodecov("within_student"))
WIBnet18.fit.511 = bergm(ntwk.511.18 ~ edges + nodecov("within_student"))
WIBnet19.fit.511 = bergm(ntwk.511.19 ~ edges + nodecov("within_student"))
WIBnet20.fit.511 = bergm(ntwk.511.20 ~ edges + nodecov("within_student"))

# Weighted ERGM ----
wghtnet.fit.511.wc <- ergm(
  ntwk_wght.item511 ~ sum + nodecov("within_action"),
  response = "nominations",
  reference = ~Poisson,
  control=control.ergm(MCMC.samplesize=10000,
                       MCMC.burnin=100000))

# BERGMs ----
Bnet511.fit1 <- btergm(ntwk.511 ~ edges+ nodecov("within_student"))
Bnet511.fit2 <- btergm(ntwk.511 ~ edges+ nodecov("across_student"))
Bnet511.fit3 <- btergm(ntwk.511 ~ edges+ nodecov("within_action"))
Bnet511.fit4 <- btergm(ntwk.511 ~ edges+ nodecov("within_student") + nodecov("across_student") +
nodecov("within_action"))
```

Appendix F

Aim 3 R Code: HMMs for Response Groups (Item 511)

```
# Run HMM for all players (6-state model was selected) ----
naction <- length(unique(long.511$Action_new))
ntime <- as.numeric(wide.511$Sequence.Length)

m.6 <- depmix(response = Action_new~1, data = long.511, nstates = 6,
              family = multinomial('identity'),respstart = runif(6*naction),
              trstart = runif(36), instart = runif(6), ntimes = ntime)
fm.6 <- fit(m.6,emc = em.control(random.start = F))
summary(fm.6)

# Visualize latent states and transitions ----
g1 <- graph_from_adjacency_matrix(matrix.correct, mode="directed", weighted=TRUE)
vertex_attr(g1)
E(g1)$width <- 1+E(g1)$weight*4 # change width of edges based on weights
plot(g1, layout=layout_as_star,
     edge.arrow.size=1, vertex.size=10, veterx.label.cex=1,
     vertex.color="lightgoldenrod",
     edge.color="lavender",
     vertex.label.dist=c(3,-2.5),
     edge.label=E(g1)$weight, edge.label.size=0.9,
     vertex.label.dist=c(2,2), edge.label.dist=7)
```

Appendix G

Aim 4 Mplus Code: 2-Class LCA Models (Item 511)

Model 1: Predictor-Outcome Relation Constrained across Classes

```
TITLE:
LCA s1-s19 completers as other code 2 class full mod Y=item perf CONSTR;
DATA:
FILE=C:\LCA_data.csv;
VARIABLE:
NAMES =
ID
TotTime
TotAtt
AttCorr
PctCorr
ItemCt
ItemResp
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19
LogTotTm
LogTotAt
LogAttCr
LogItmCt
;
USEVARIABLES =
ID
LogTotTm
LogItmCt
ItemResp
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19
;
```

```
IDVARIABLE = ID;
CLASSES = c(2);
CATEGORICAL = ItemResp;
NOMINAL =
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19;

SAVEDATA:
FILE = LCA_2class_10completers_fullmodel_1.dat;
SAVE = cprobabilities;

ANALYSIS:
TYPE = mixture;

MODEL:
%OVERALL%
c ON LogTotTm;
LogItmCt ON LogTotTm;
ItemResp ON LogTotTm;

OUTPUT:
stdyx tech11 tech14;
```

Model 2: Predictor-Outcome Relation Unconstrained across Classes

```
TITLE:
LCA s1-s19 completers as other code 2 class full mod Y=engagement UNCONSTR;
DATA:
FILE=C:\LCA_data.csv;
VARIABLE:
NAMES =
ID
TotTime
TotAtt
AttCorr
PctCorr
ItemCt
ItemResp
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19
LogTotTm
LogTotAt
LogAttCr
LogItmCt
;
USEVARIABLES =
ID
LogTotTm
LogItmCt
ItemResp
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19
;

IDVARIABLE = ID;
CLASSES = c(2);
CATEGORICAL = ItemResp;
```

```
NOMINAL =
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19;

SAVEDATA:
FILE = LCA_2class_10completers_fullmodel_1.dat;
SAVE = cprobabilities;

ANALYSIS:
TYPE = mixture;

MODEL:
%OVERALL%
c ON LogTotTm;

%C#1%
LogItmCt ON LogTotTm;
ItemResp ON LogTotTm;

%C#2%
LogItmCt ON LogTotTm;
ItemResp ON LogTotTm;
```

```
NOMINAL =
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19;

SAVEDATA:
FILE = LCA_2class_10completers_fullmodel_2.dat;
SAVE = cprobabilities;

ANALYSIS:
TYPE = mixture;

MODEL:
%OVERALL%
c ON LogTotTm;
LogItmCt ON LogTotTm;
ItemResp ON LogTotTm;

%C#1%
LogItmCt ON LogTotTm;
ItemResp ON LogTotTm;

%C#2%
LogItmCt ON LogTotTm;
ItemResp ON LogTotTm;

OUTPUT:
stdyx tech11 tech14;
```