Identifying Cognitive and Meta-cognitive Aspects of Digital Inquiry from Process Data

Jesse R. Sparks & Caitlin Tenison

Educational Testing Service

Beyond Results Workshop – October 1, 2021

NAEP SAIL Initiative

The Survey Assessment Innovations Laboratory (SAIL) is a hothouse dedicated to explorations in the cognitive sciences, assessment, and technology that will enable a continuously updated NAEP which leads the field of educational assessment

New models of 21st century skills and competencies

New technologies to elicit and capture complex human interactions

Computational advances for automated scoring, complex modeling, score reporting

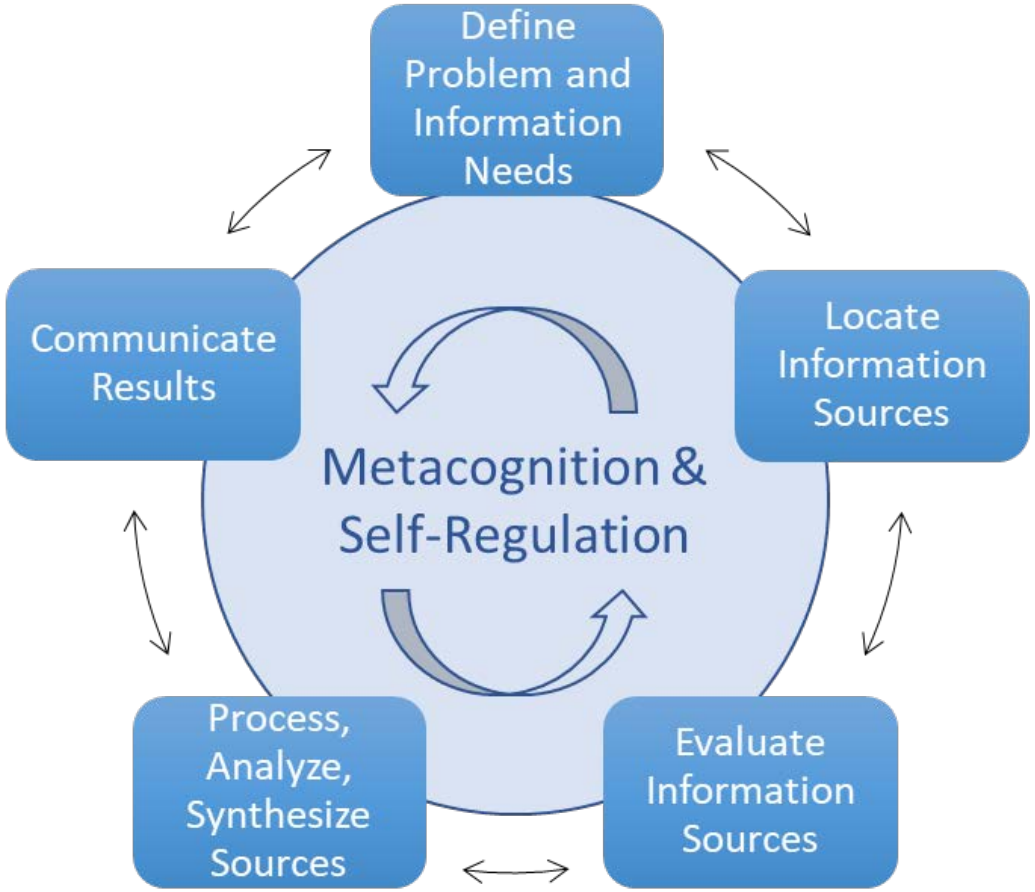
NAEP SAIL Virtual World Project – Motivation

- There is a growing interest in teaching and learning of complex constructs:
  - Conducting **inquiry from multiple sources** in the English Language Arts – a type of multiple-text comprehension or multiple-source use
- There is a **need to develop quality large-scale assessments** that provide valid information about students' knowledge, skills, and abilities as related to their processes and performances within authentic, complex inquiry tasks.
  - Allow us to capture dynamic inquiry processes, vs. static knowledge measures
- There is a **need for correspondingly complex analytic approaches** to evaluate the processes and performances captured in more authentic, complex inquiry tasks.
  - Models should account for the interplay between individuals' KSAs, their goals, the task goals, and task design features

NAEP SAIL Virtual World Project – Aims

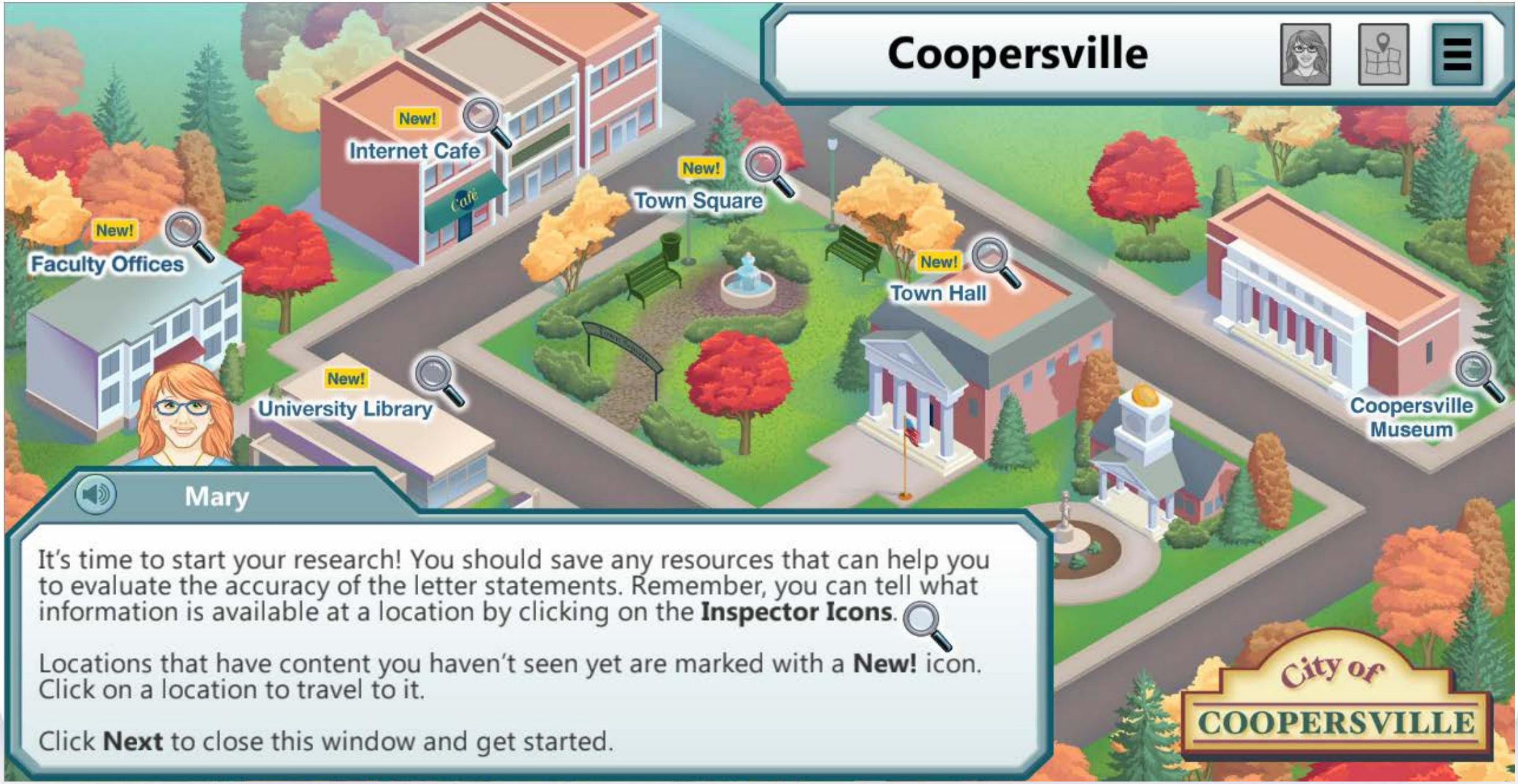
- Research and develop **virtual environments** with authentic contexts
  - To gather evidence of students' inquiry skills **using naturalistic simulations** reflecting real-world applications of the targeted skills
- Explore students' interactions in inquiry scenarios that require
  - **Locating, evaluating, reading**, and **writing syntheses** from multiple sources
  - Reconciling conflicting, unreliable, and inaccurate information
- Leverage technology from games and simulations
  - to **enhance student motivation and engagement** in assessment tasks
  - to **provide evidence of students' cognitive processes** as they unfold in time, allowing application of new analytic approaches
    - (i.e., designed to support analysis of process data captured in log files).

SAIL ELA Construct: Digital Multiple-Source Inquiry

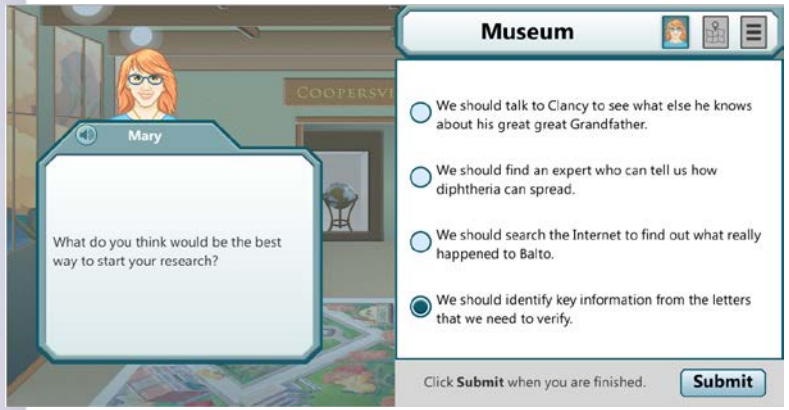


- **Inquiry from multiple sources in history and science**
  (e.g., Britt & Aglinskas, 2002; Britt & Rouet, 2012; Goldman et al., 2010, 2011, 2013; Graesser et al., 2007; Perfetti et al., 1999; Rouet & Britt, 2011; Wiley et al., 2009; Wineburg, 1991, 1998)
- **Critical evaluation of multiple sources**
  (e.g., Braasch et al., 2009, 2012; Bråten et al., 2009; Graesser et al., 2007; Sparks, 2013; Sparks & Deane, 2015; Strømsø et al., 2013; Wiley et al., 2009)
- **Online reading comprehension and theory of "new literacies"**
  (e.g., Coiro & Dobler, 2007; Coiro & Kennedy, 2011; Coiro et al. 2013; Leu et al., 2008, 2015)
- **Information problem solving and inquiry in digital environments**
  (e.g., Baker & Clarke-Midura, 2013; de Jong, 2006; Linn et al., 2003; Kuiper et al., 2005; Quintana et al., 2004; Wallace et al., 2000; Walraven et al., 2008; Zhang & Quintana, 2012)
- **Assessments of related constructs**
  (e.g., ETS CBAL assessments, Reading for Understanding initiative Global Integrated Scenario-Based Assessments (GISA), CWRA+/CLA+, Online Research and Comprehension Assessment, ePIRLS, Project READi multiple-source comprehension assessments, ETS iSkills digital literacy assessment)

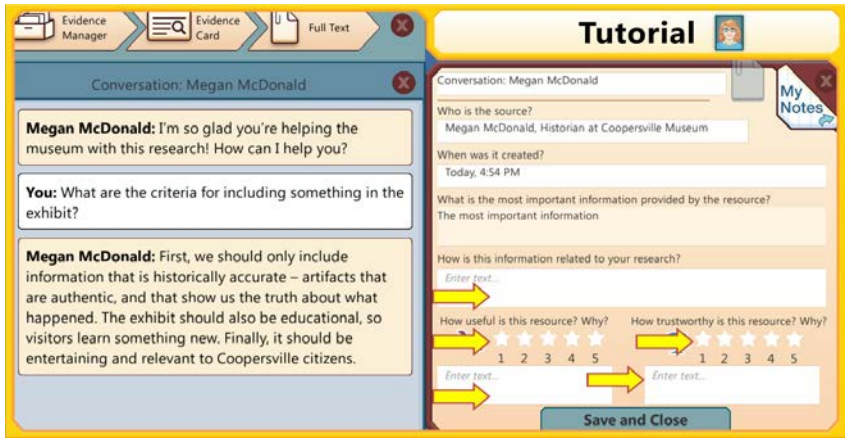
SAIL ELA Virtual World Environment



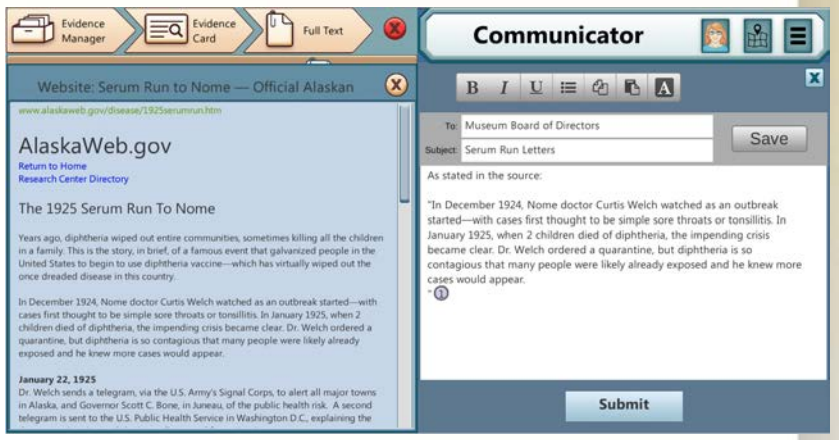
SAIL ELA Virtual World: Tools and Resources



Planning Inquiry with Virtual Partner (Guide)



Supports for Source Evaluation & Note-taking for Saved Resources



Compose Written Response to Inquiry Task using Collected Resources

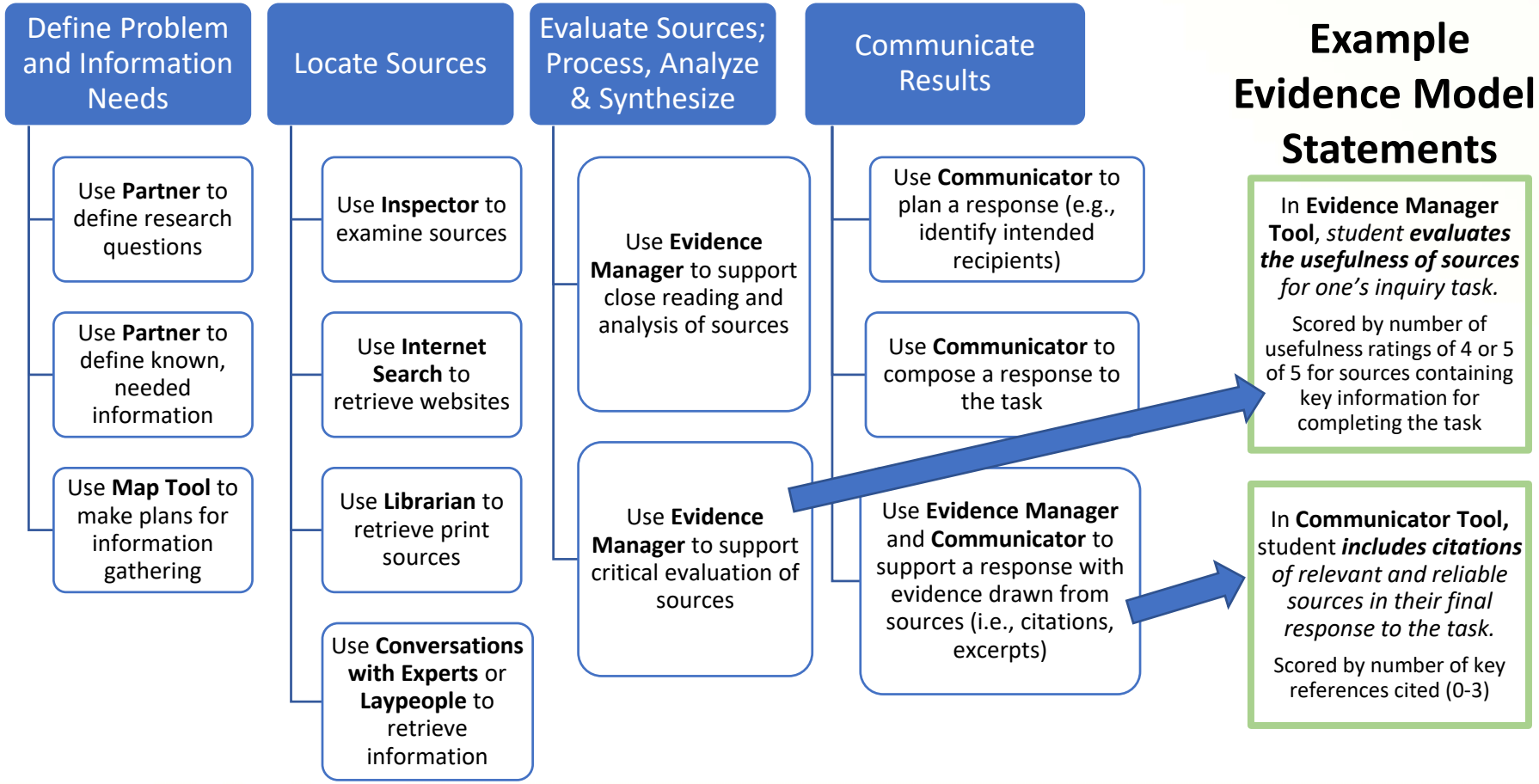
Simulated Web Search & Library Search Tools to retrieve documents

Simulated Questioning of Virtual Characters (experts or laypeople)

SAIL ELA Student/Evidence Model Alignment

Each aspect of the **SAIL ELA Virtual World** is designed to measure a specific aspect of the **online inquiry** construct. All student actions are mapped to a specific dimension, and are evaluated with specific scoring rules. In ECD parlance, this represents our **Evidence Model**.

Relations between Online Inquiry Construct & SAIL ELA Virtual World digital tools (white boxes):



Example Evidence Model Statements

In **Evidence Manager Tool**, *student **evaluates the usefulness of sources** for one's inquiry task.*

Scored by number of usefulness ratings of 4 or 5 of 5 for sources containing key information for completing the task

In **Communicator Tool,** student ***includes citations** of relevant and reliable sources in their final response to the task.*

Scored by number of key references cited (0-3)

SAIL ELA Scenario-Based Task Structure

- **Goal**: Evaluate historical accuracy of claims contained in artifact—Should it go in museum?

- Three phases:
  - **Setup**,
  - **Free Roam**, and
  - **Conclusion**

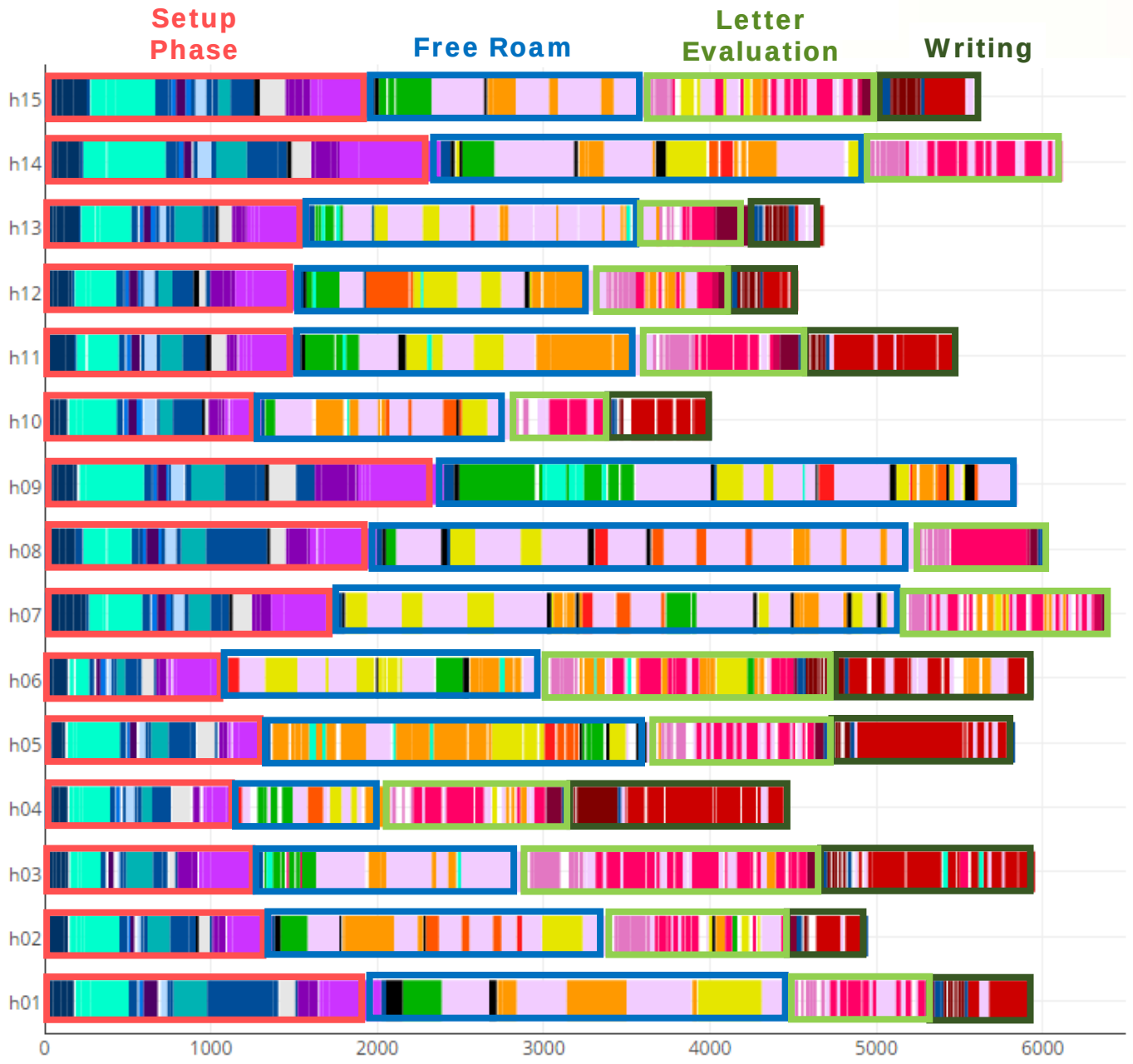
  - Free Roam allows for **nonlinear, free exploration** and investigation of **inquiry process**.

Phase	Construct(s)	Points (100 total)
Setup: *Planning Inquiry*	Define Problem/Inform. Needs, Locate, Evaluate	**12** 1 best first step 3 claims to examine 2 useful locations 1 question to authority figure 5 Evidence Manager evaluations
Free Roam: *Gather and Evaluate Resources*	Locate, Evaluate, Process/Analyze/ Synthesize	**51** 8 *Library* (2 search, 1 save, 5 EM Evaluations) 20 *Faculty Offices* (6 questions, 2 talk, 2 save, 10 EM Evaluations) 23 *Internet Café* (2 search, 3 view, 3 save, 15 EM evaluations)
Conclusion: *Evaluation Task*	Evaluate, Process/Analyze/ Synthesize	**27** 10 true/false judgments 10 source attributions 6 corrections 1 final decision on the overall inquiry task
Conclusion: *Argument Writing Task*	Process/Analyze/ Synthesize, Communicate	**10** 3 citations 1 excerpt 6 argument quality (NAEP Writing to Persuade)

Cognitive Lab Study Data (N=15)

Time per Scene/ Tool

each row is one participant



Setup Phase
IntroScenario
Instructions
ReadLetters
Tutorial_Inspector
Partner
SetupQ
ChooseStatementsQ
ChooseLocationsQ
Tutorial_Conversation
Tutorial_EvMan
Free Roam Phase
Map
FacultyOffices
Library
Toogle
TownHall
Townsquare
EvidenceManager
Conclusion Phase
Tutorial_LetterEvaluation
LetterEvaluation
FinalDecision
LetterEvaluation_Correction
Tutorial_Communicator
Communicator

Virtual World Tryout Study – Sample & Dataset

N=130 8th Graders

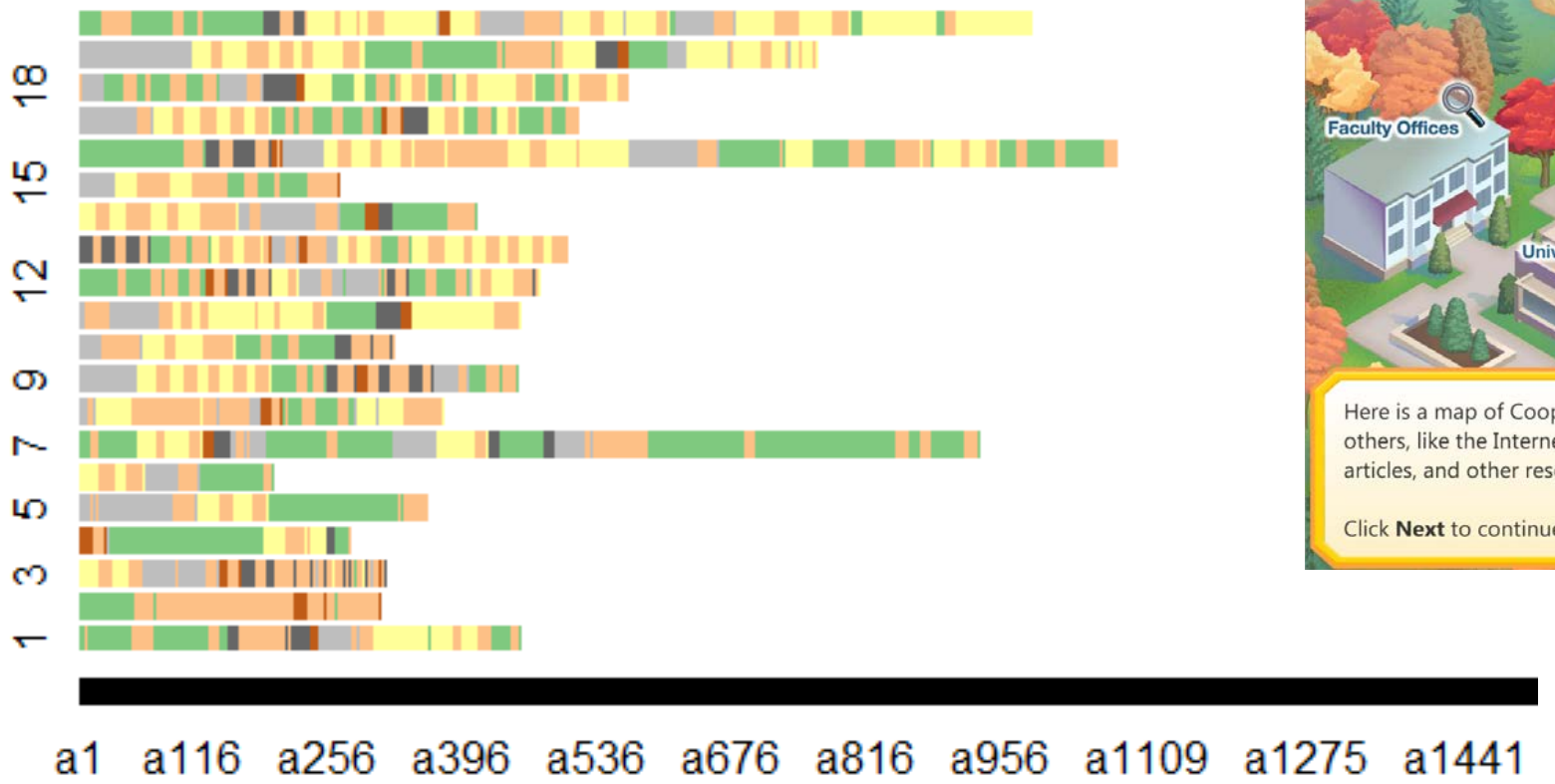
- 91 (70%) from Urban NJ
- 37 (30%) from Rural AR

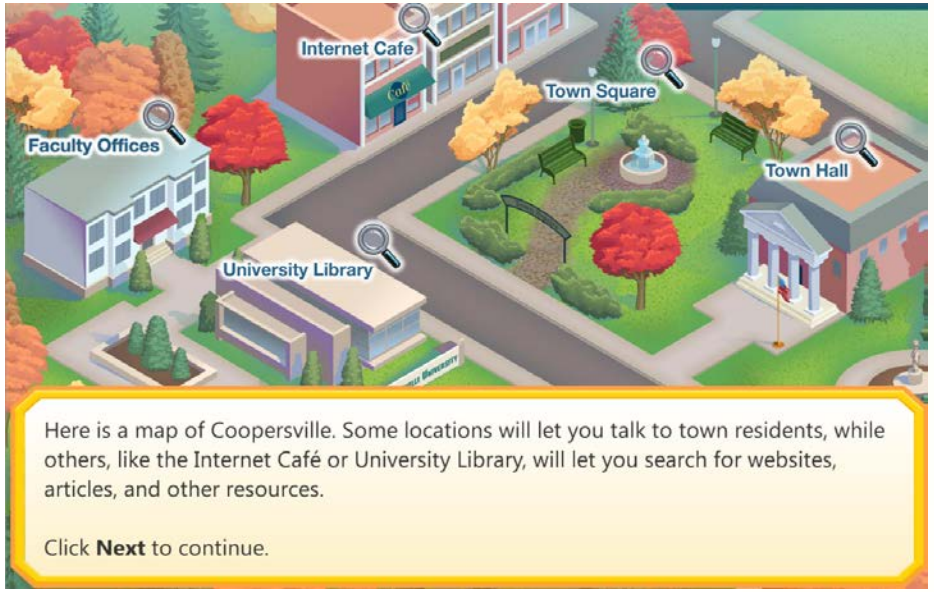
- 3 missing roster data
- Some students had multiple sessions (errored, then restarted from the beginning)
  - "most complete" session (with most scoring opportunities) was retained for scoring & analysis.

- 32 (~25%) did not complete the task through the culminating Essay task.

- 47 (37%) white students,
  80 (63%) students of color:
  - 25 Asian/Asian American
  - 11 Black/African American
  - 44 Hispanic/Latino
- 67 (53%) female,
  60 (47%) male
- 92 (72%) participate in FRPL
- 6 received Title I accommodations for state tests.

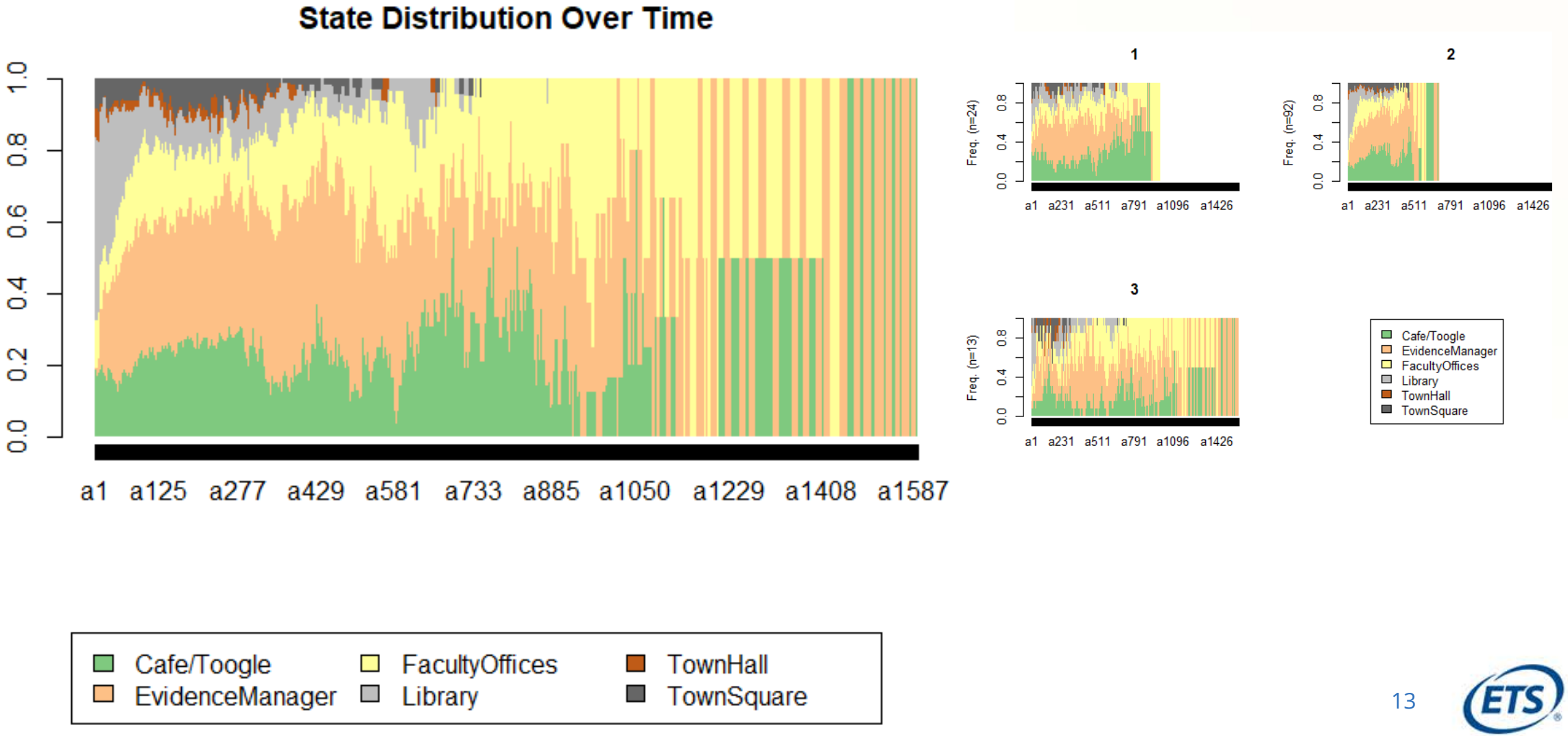
Tryout Study - 20 Sample Free Roam Sequences



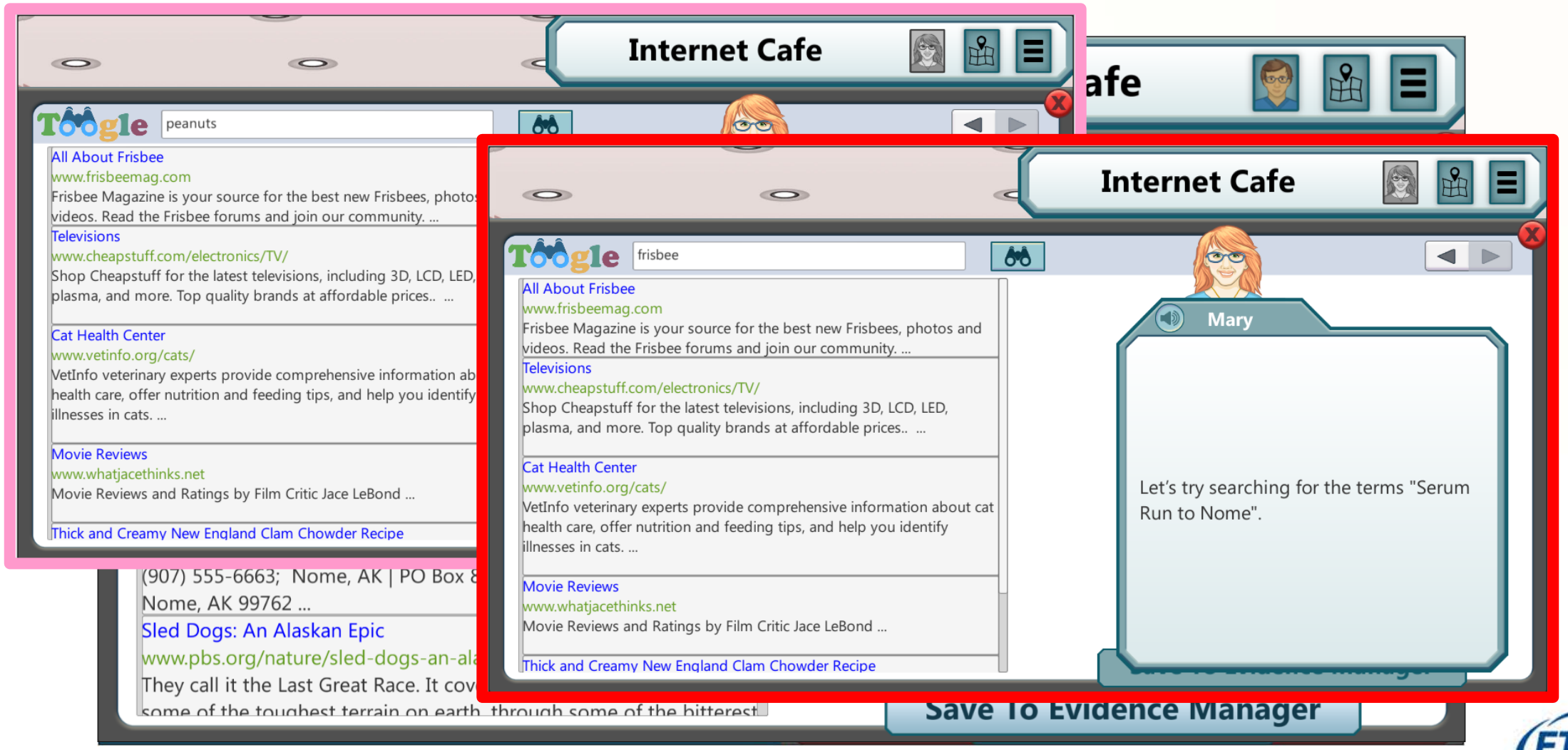


Cafe/Toogle
EvidenceManager
FacultyOffices
Library
TownHall
TownSquare

Tryout Study - State Distribution Analysis



Simulated Web Search



Modeling Student Search Strategies

- Think-aloud studies of web searches document various web search strategies and heuristics used by participants at different points throughout the task.

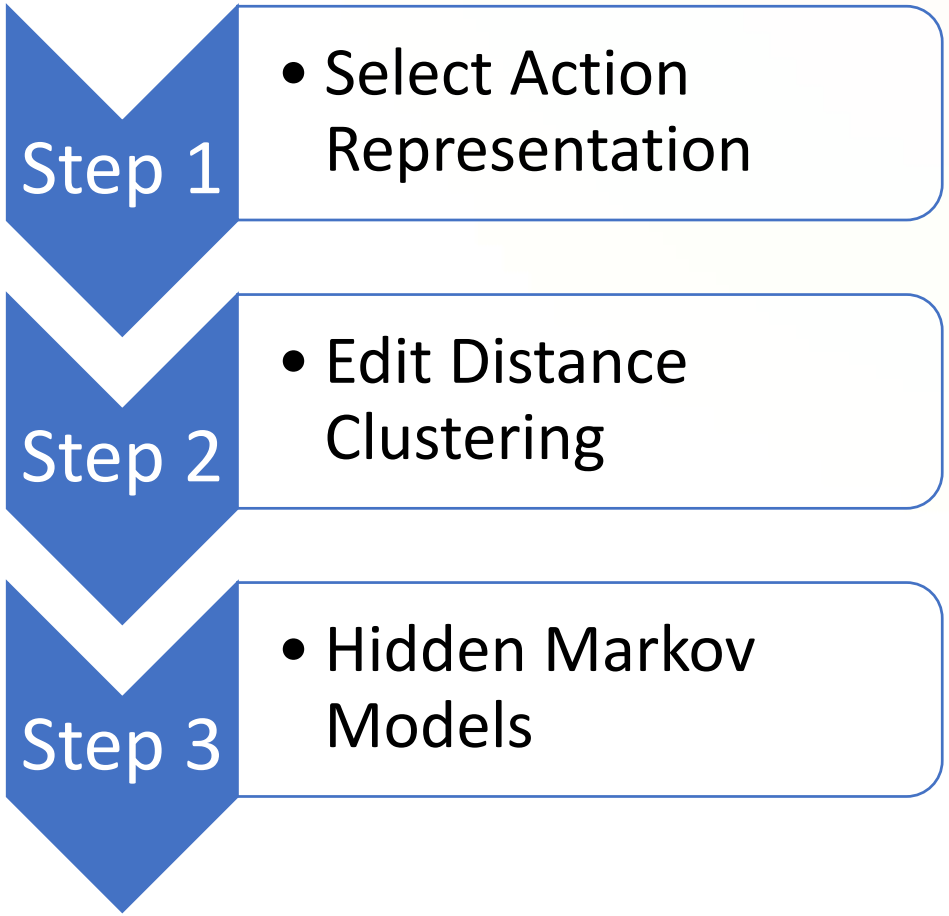
- Prior research in cognitive science suggests that search strategies reflect individuals' KSAs, as well as their goals, the task goals, and task design features

- Can we detect **meaningful** differences in student's strategy use?
  - Reflect different cognitive processing
  - Lead to different outcomes
  - Correlate with different overall task performance

Modeling Student Search Strategies

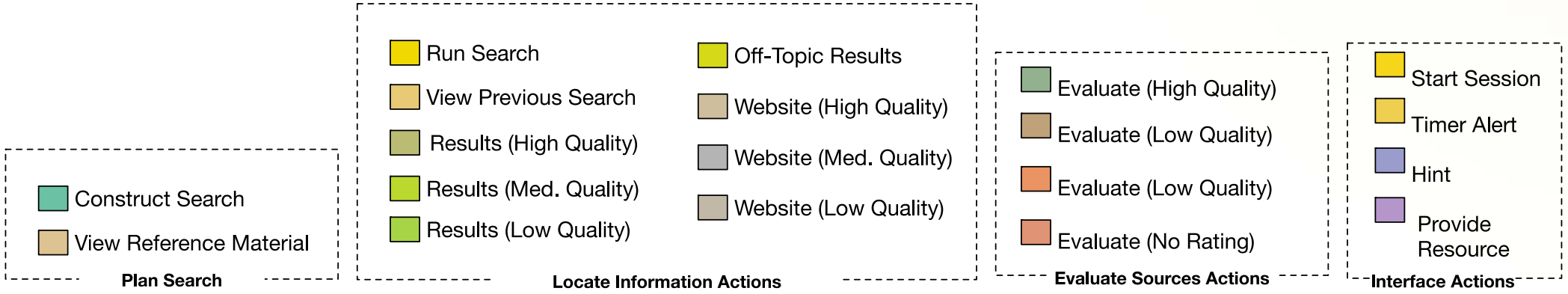
- Our aim in clustering students' search processes is to identify distinct patterns that help us characterize students' ability to plan their search, locate websites, and evaluate those websites.

- We used a multi-step clustering method to account for both the **context** and **content** of students' actions.



Step 1: Action Representation

Model descriptiveness is directly affected by how we represent students' search processes



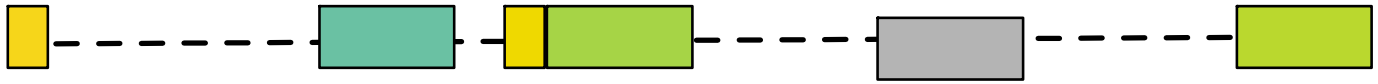
- **Actions**:
  - Student actions that reflect construct relevant activities: planning, locating information, and evaluating sources.
  - Interface actions/changes in task state
  - Provide information about the quality of the action
  - Standardize our sampling of task execution (e.g. don't over sample some tasks and under sample others)

Step 1: Action Representation

Pauses:

- Identify pauses that reflect significant cognitive processing
- Provide rough distinction between types of processes (e.g. plan execution, encoding information, planning, etc)

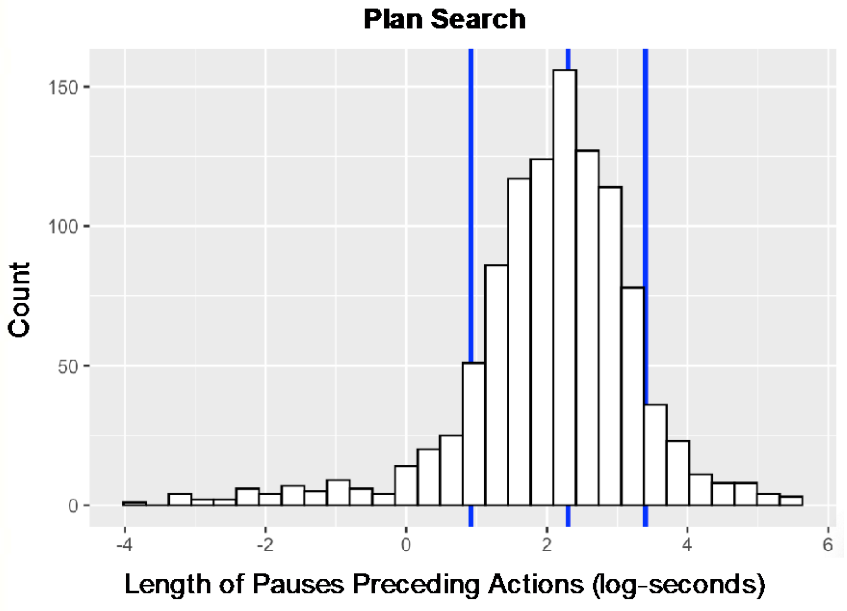
Timing of the actions

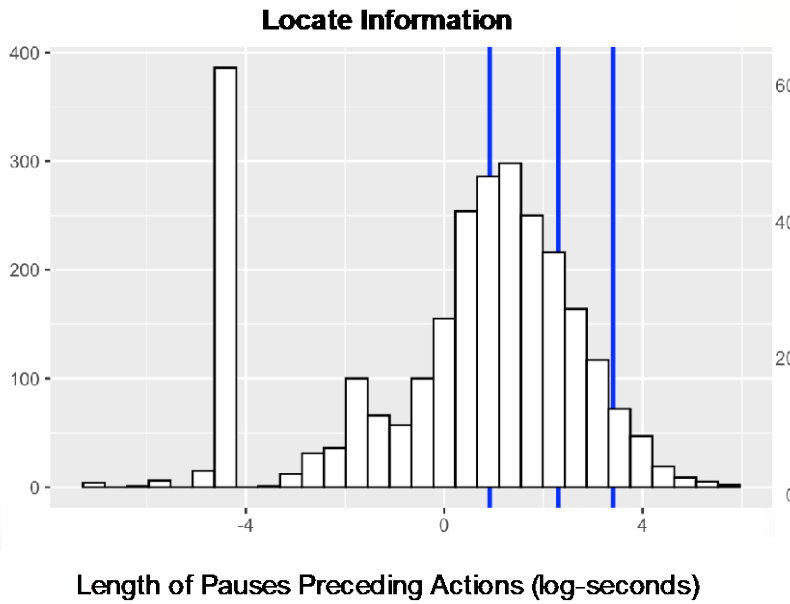


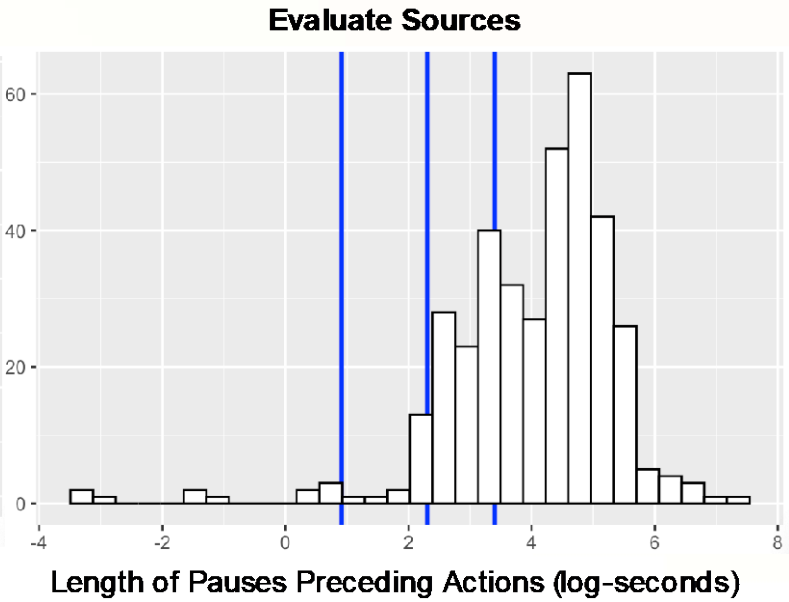
Simple HMM Representation of actions does not capture timing

Adding representation of 'Pause categories'

Step 1: Action Representation

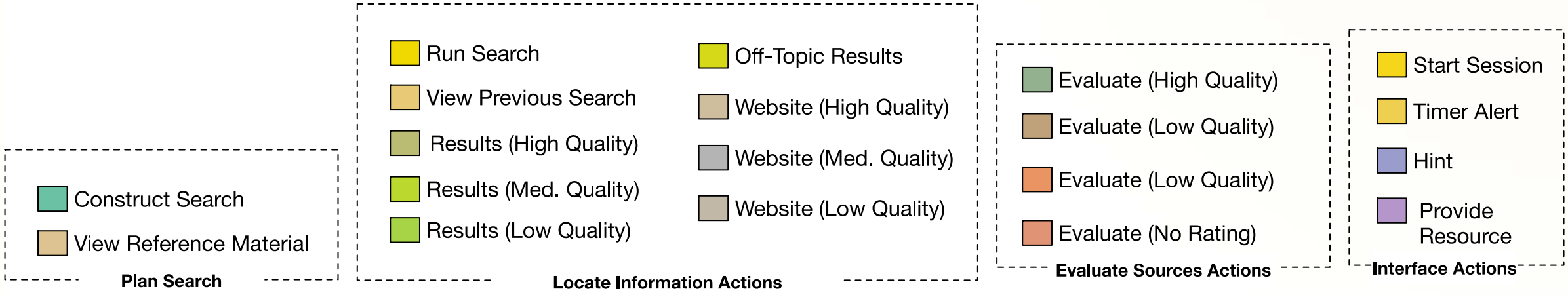






	Description	Label	Freq.
Pauses between actions	Between 2.5 and 10 seconds	Short Pause	19.9% (8.2%)
	Between 10 and 30 seconds	Medium Pause	12.8% (6.7%)
	Greater than 30 seconds	Long Pause	9.3% (6.6%)

Step 1: Action Representation



Example Sequence

Start Session	Long Pause	Construct S.	Short Pause	Run S.	Results Med	Med. Pause	Website Med

Step 2: Edit distance clustering

Goal:

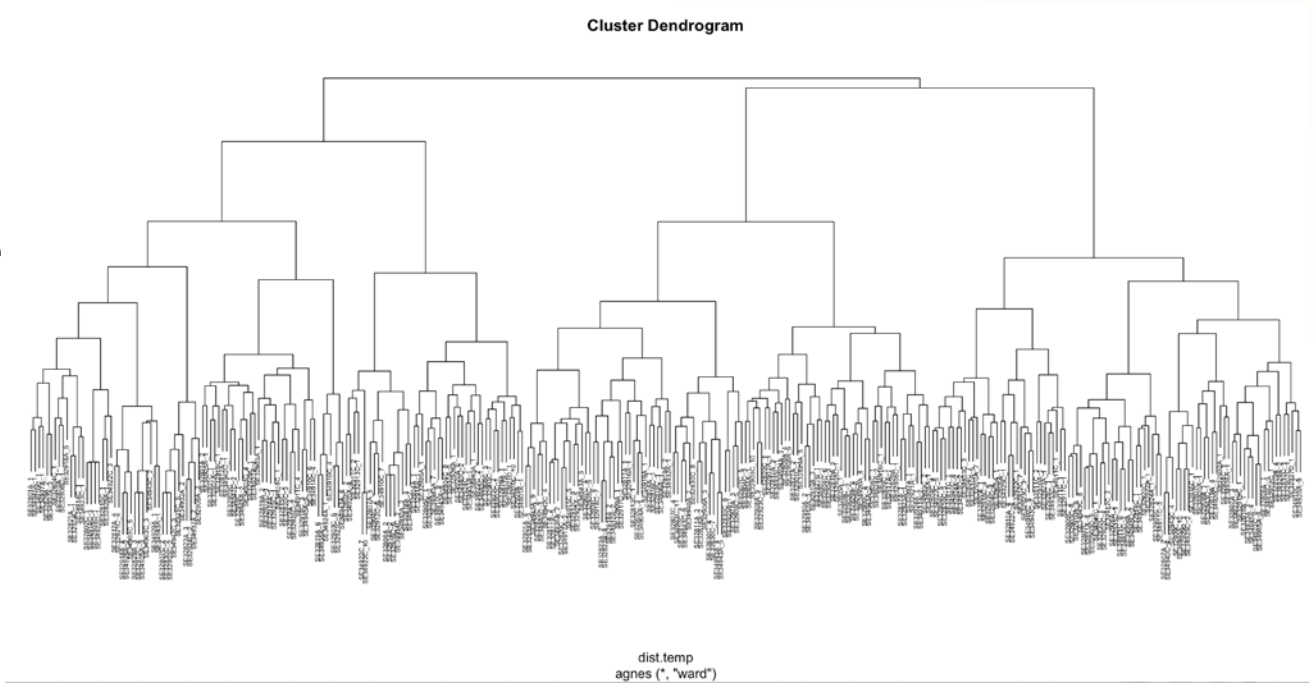
- Identify similar sequences

Approach:

- We used a normalized *optimal matching* (OM) metric to calculate the distance between all sequence.
- We then applied hierarchical agglomerative clustering on the resulting pairwise OM distances to identify clusters of similar search sequences

Result:

- We use this to identify 4 cluster categories



Step 3: Hidden Markov Models

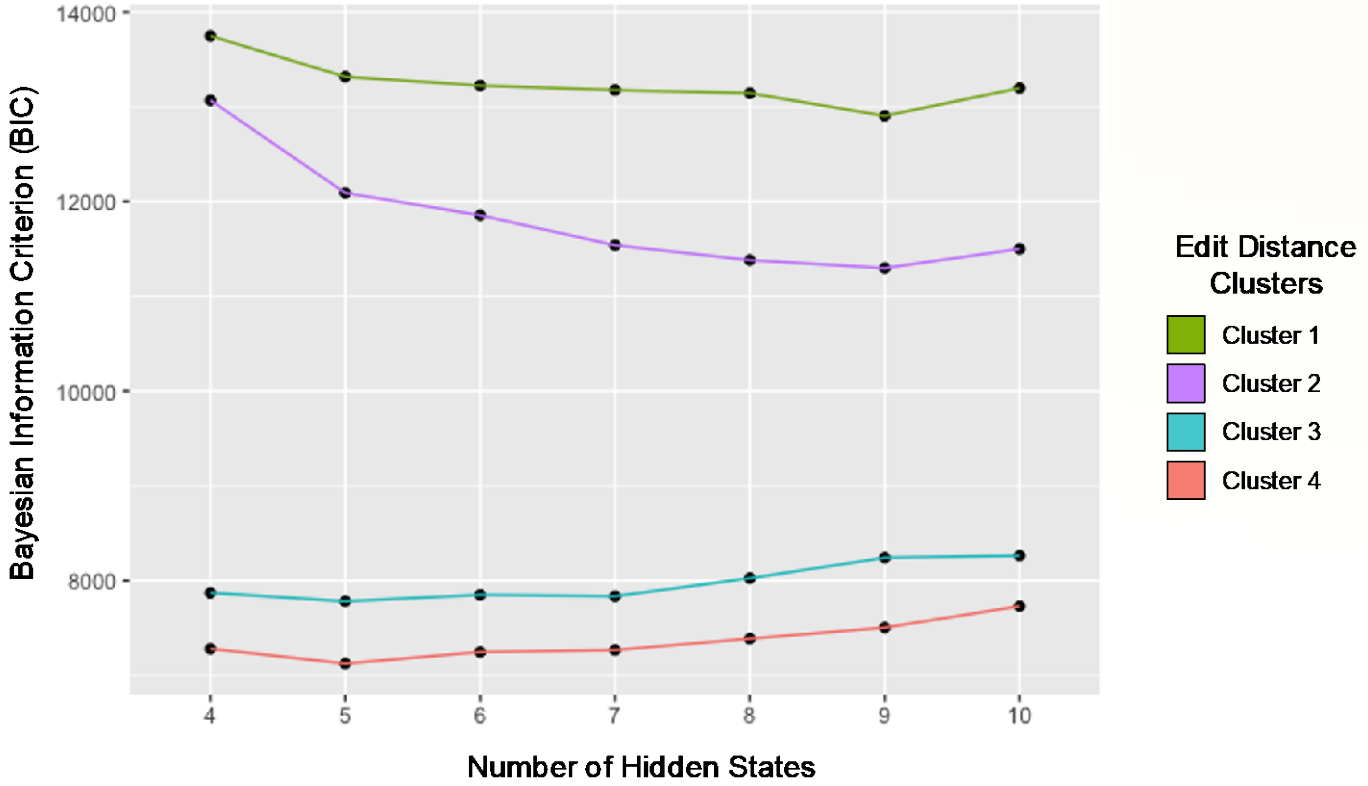
While Edit distance clustering is a well-established method for clustering educational process data (Hao et al., 2015; Boroujeni & Dillenbourg, 2019), it **does not** consider the order in which actions occur, which can complicate interpretation.

Hidden Markov model (HMMs) capture the context of actions by modeling the probabilistic transition between latent action states.

- These latent action states reflect a latent state that incorporates both the content and context of the action.
- We grouped sequences in terms of four edit-distance clusters and fit four separate HMMs, one for each cluster.

Step 3: Hidden Markov Models

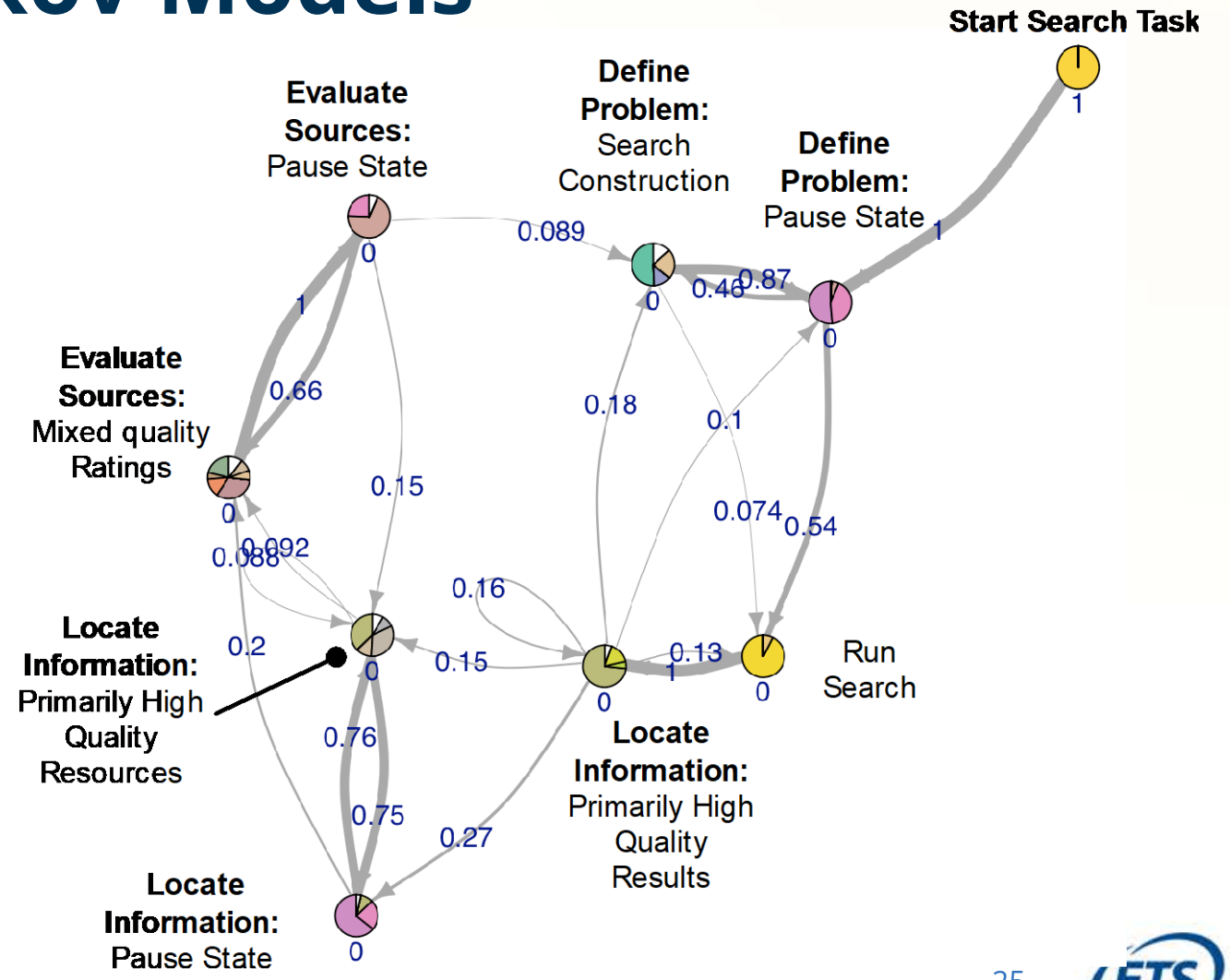
- Bayesian Information Criteria (BIC) indicates goodness of fit while penalizing for added parameters.
- Lower scores indicate better fits.
- Cluster 1 and Cluster 2 sequences are best fit by **9 state models.**
- Cluster 3 and Cluster 4 sequences are best fit by **5 state models**.



Step 3: Hidden Markov Models

Cluster 1:

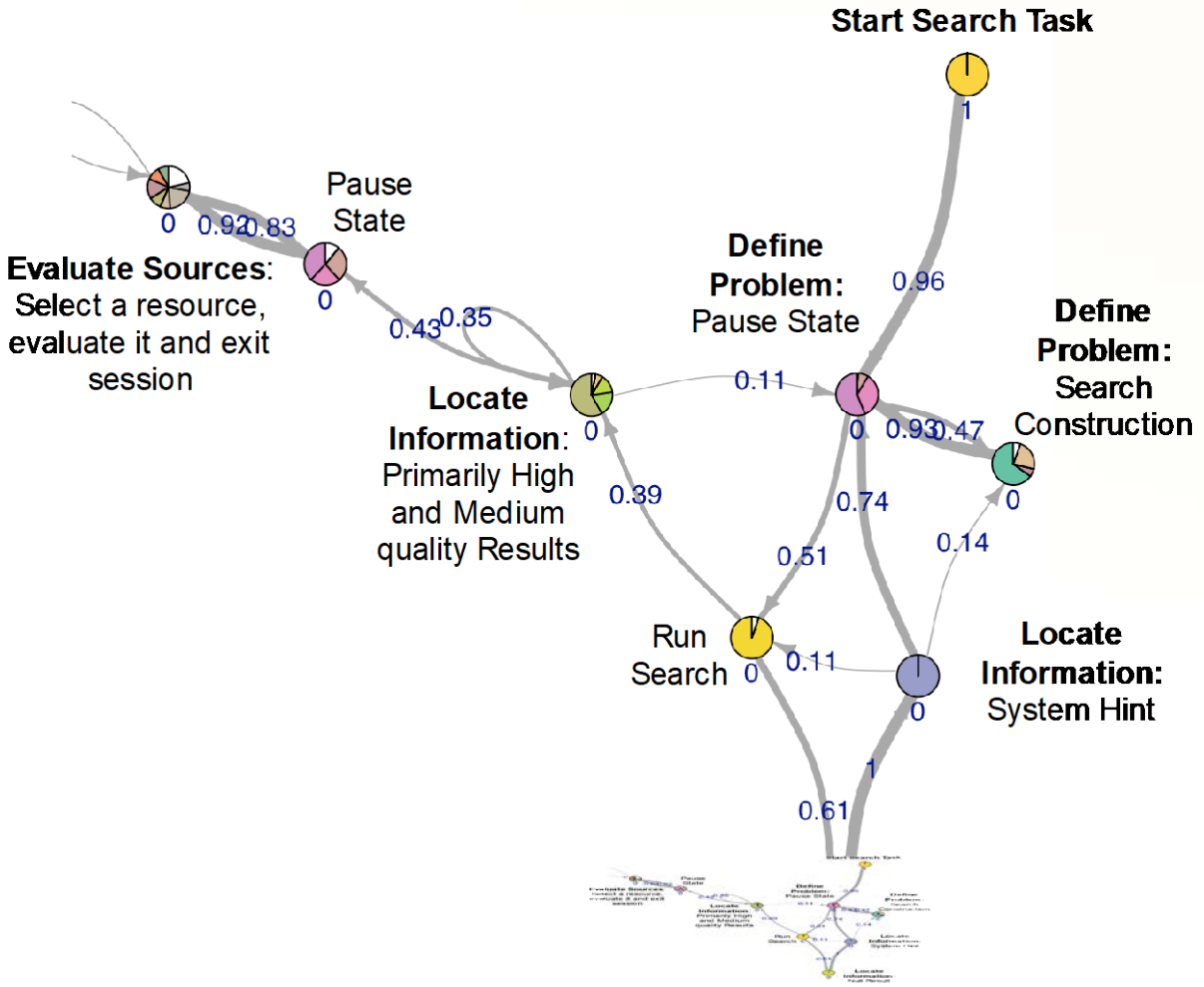
- n=102 sequences, 26.8% of first sessions
- appears to reflect a relatively proficient strategy use.
- Searches had a high probability of yielding high-quality results, which contain the two most useful websites within the task



Step 3: Hidden Markov Models

Cluster 2:

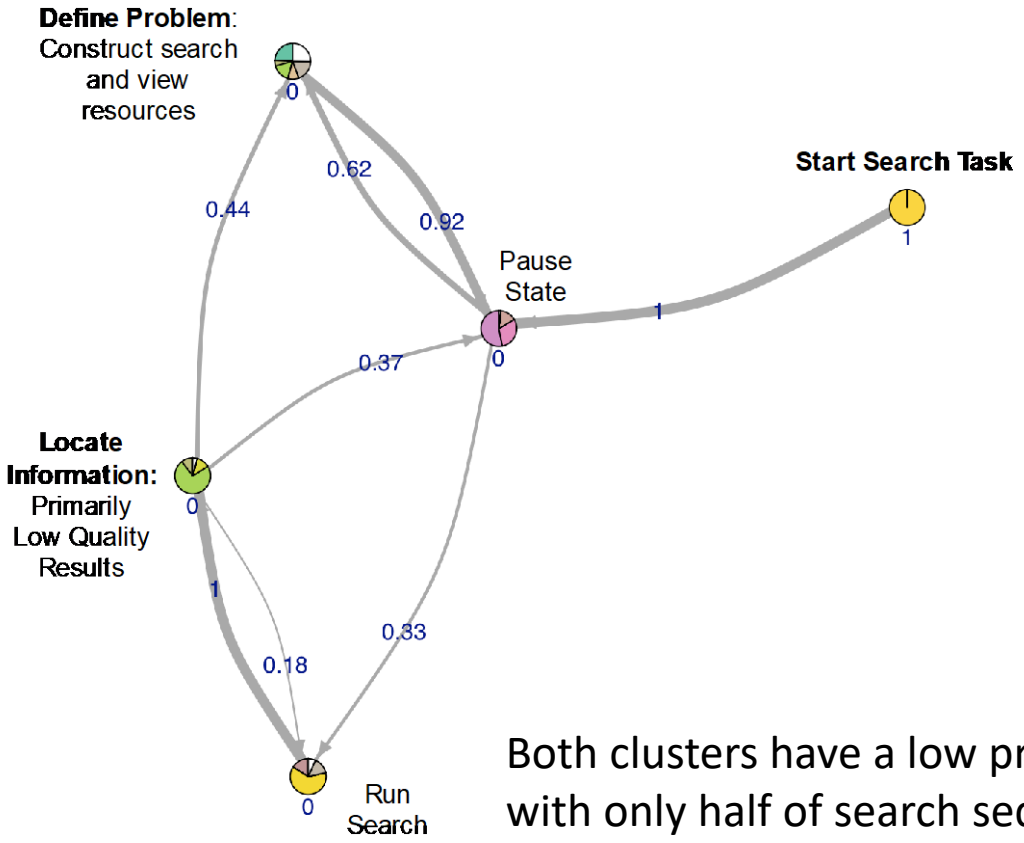
- n=93 sequences, 56.9% of first sessions
- Characterized primarily by the presence of an off-topic search yielding off-topic results
- This strategy captures unsuccessful searches resulting in system hints
- Unlike Cluster 1, this strategy features a more direct path from the results page to the selection of a website and completion of the resource evaluation prompts



Step 3: Hidden Markov Models

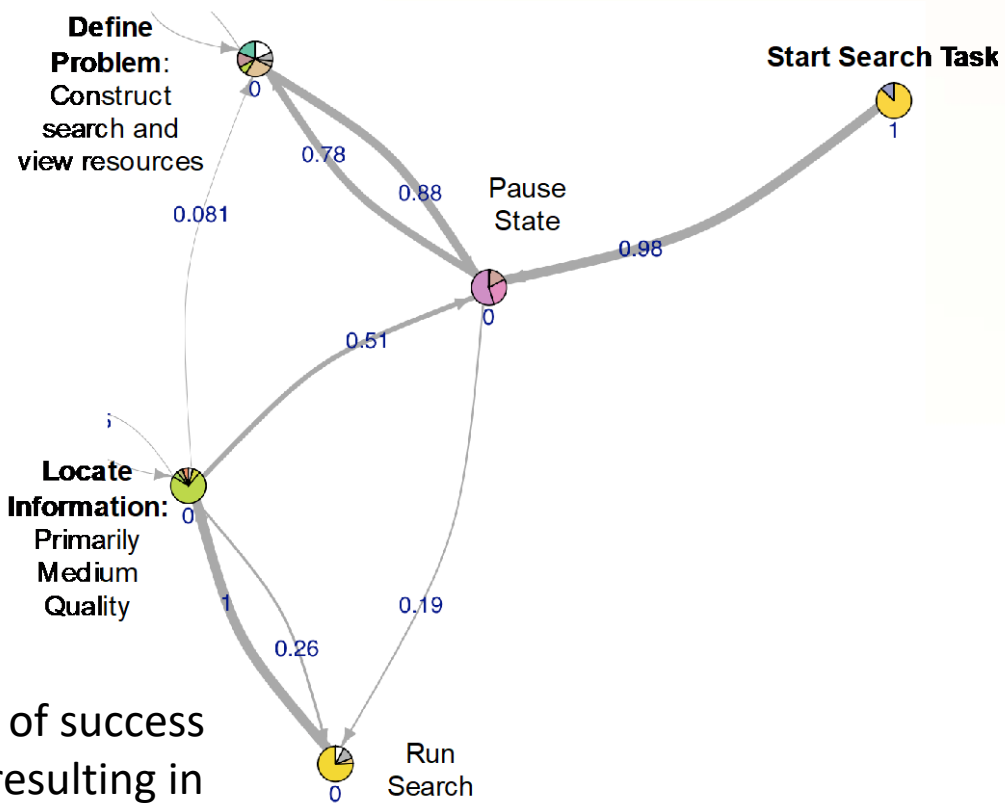
Cluster 3:

n=79 sequences, 11.4% of first sessions



Cluster 4:

n=45 sequences, 4.9% of first sessions



Both clusters have a low probability of success with only half of search sequences resulting in saving and evaluating a website in the Evidence Manager (Cluster 3: 42%, Cluster 4: 50%).

Relationships Between Strategies and Scores

Spearman correlations

	Proportion Sessions in **Cluster 1**	Proportion Sessions in **Cluster 2**	Proportion Sessions in **Cluster 3**	Proportion Sessions in **Cluster 4**
Total Task Score				
Inquiry Task Total Score (max: 100)	**.310**	**-.325**	.035	.019
Task Phase-Level Subscores				
Task Phase: Setup (max: 12)	**.228**	**-.251**	.066	-.017
Task Phase: Free Roam (max: 51)	**.209**	**-.260**	.044	.098
Task Phase: Conclusion (max: 37)	**.203**	**-.259**	.117	-.005
Construct Subscores				
Subconstruct: Planning (max: 6)	.100	**-.240**	**.207**	.083
Subconstruct: Locating (max: 22)	.150	-.197	.084	.015
Locating: Questioning (max: 7)	-.053	-.089	**.202**	.083
Locating: Searching (max: 4)	**.407**	**-.254**	**-.226**	-.041
Locating: Choosing Sources (max: 5)	.095	-.112	.070	-.053
Locating: Saving Sources (max: 6)	.117	-.176	.112	.001
Subconstruct: Evaluating (max: 35)	**.260**	**-.275**	-.019	.099
Evaluating: Importance (max: 7)	**.215**	**-.257**	.029	.095
Evaluating: Usefulness (max: 14)	**.247**	**-.236**	-.038	.052
Evaluating: Trustworthiness (max: 14)	**.225**	**-.245**	-.025	.118
Subconstruct: Synthesis (max: 37)	**.203**	**-.259**	.117	-.005

Discussion – SAIL ELA Full Task

- Virtual world task and tools **capture differences in inquiry processes**
  - Many different paths and time allocations across locations and resources
- Process data produced at the full task level was challenging to map to inquiry cycle
  - Future work may benefit from a more careful selection of events to reflect critical processing or alternative modeling methods

Discussion

Aim: Identify meaningful distinctions in inquiry strategies that incorporates the **quality** of the materials interacted with, the **time** spent on different actions, and the **context** of those actions

Do we see **meaningful** differences in student's strategy use?

- Reflect different cognitive processing
  - We see differences in the length and contexts in which learners pause for clusters 1 and 2 suggesting some sensitivity
  - These differences are not present in cluster 3 and 4 – unclear
- Lead to different outcomes
  - We see differences in student's searches ending in the selection and saving of task relevant resources.
  - Future work should consider how these strategies are situated within the larger task context
- Correlate with different overall task performance
  - We see trends however as one of many smaller subtasks within the full virtual world these correlations were quite small.

Discussion

The Method: there are many choices which that combine an intuition for the construct of interest, understanding of human cognition, and insight into the behavior of the modeling approaches.

- Future work will need to identify means for standardizing and guiding these choices.
- How could we scale this to redesign tasks and replicate results

The Evaluation: How do we evaluate when an approach like this is successful?

- Is linking process to these outcome measures the gold-standard and should it be?
  - This is construct dependent
- Are there other features we could engineer within the task or modify about the design of the instrument to make this a better proxy/ more closely related?
  - Changing what we log
  - Introduce task constraints
  - Consider how hints and help should impact outcome measures

To follow up with questions or comments, please contact:

jsparks@ets.org or ctenison@ets.org

Acknowledgements

ELA Virtual World Project: Brian Young, Madeline Keehner, Jie Gao, Mengxiao Zhu, Colleen Appel, Gary Feng, Hilary Persky, Heather Nadelman, Irv Katz, Jonathan Steinberg, Jen Bailey, and Julie Coiro (University of Rhode Island)

Virtual World Developer: Intelligent Automation, Inc.