

# Process Data for Better Psychometrics

Lessons Learned from Response Time and Clickstream Data

Okan Bulut

Measurement, Evaluation, and Data Science

Faculty of Education, University of Alberta



bulut@ualberta.ca



www.okanbulut.com



@drokanbulut

# Outline

1. Incorporating response times into psychometrics
2. Modeling clickstream data
3. Future research directions

# Assessments → Psychometric Data



Item, testlet, and distractor position effects (e.g., Bulut, Guo, & Gierl, 2017; Bulut, Lei, & Guo, 2018; Shin, Bulut, & Gierl, 2020)

Comparing answer-until-correct and full-credit scoring in alternate assessments (e.g., Bulut et al., under review)

Effects of item format, cognitive domain, and linguistic features (e.g., Kan, Bulut, & Cormier, 2018; Liou & Bulut, 2020)

# Digital Assessments → Process Data



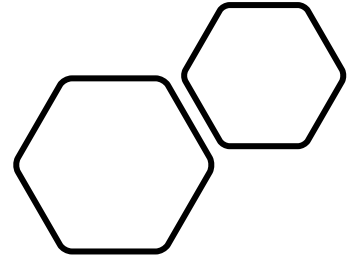
Response times

Time-stamped sequential actions (logs)

Clickstream data

Eye tracking data

Sensor data (e.g., facial expression sensor, conductance bracelet, pressure mouse, and posture analysis seat; see Arroyo et al., 2010)



**If distortion > 0**, test fraud, real-time cheating, item pre-knowledge

$$\text{Distortion} = \left[ \begin{array}{c} \textit{What the test score} \\ \textit{indicates that student} \\ \textit{i knows and can do} \end{array} \right] - \left[ \begin{array}{c} \textit{What student i} \\ \textit{actually knows and} \\ \textit{can do} \end{array} \right]$$

**If distortion < 0**, construct-irrelevant factors such as lack of test-taking motivation, disengaged responding, any type of distraction during test administration

## If distortion > 0





- Detection of compromised items using response times (Choe, Zhang, & Chang, 2018)
- Test fraud detection based on differential test-taking effort (Sinharay, 2021; Wise, Ma, & Theaker, 2014)
- Detecting examinees with item preknowledge in large-scale testing using extreme gradient (Sinharay, 2020; Zopluoglu, 2019)

## If distortion < 0

- Rapid guessing rates across administration mode and test setting (Krohne, Deribo, & Goldhammer, 2020)
- Modeling student test-taking motivation in the context of adaptive test (Wise & Kingsbury, 2016)
- Modeling examinee engagement in terms of guessing and item-level non-response (Ulitzsch, von Davier, & Pohl, 2020)

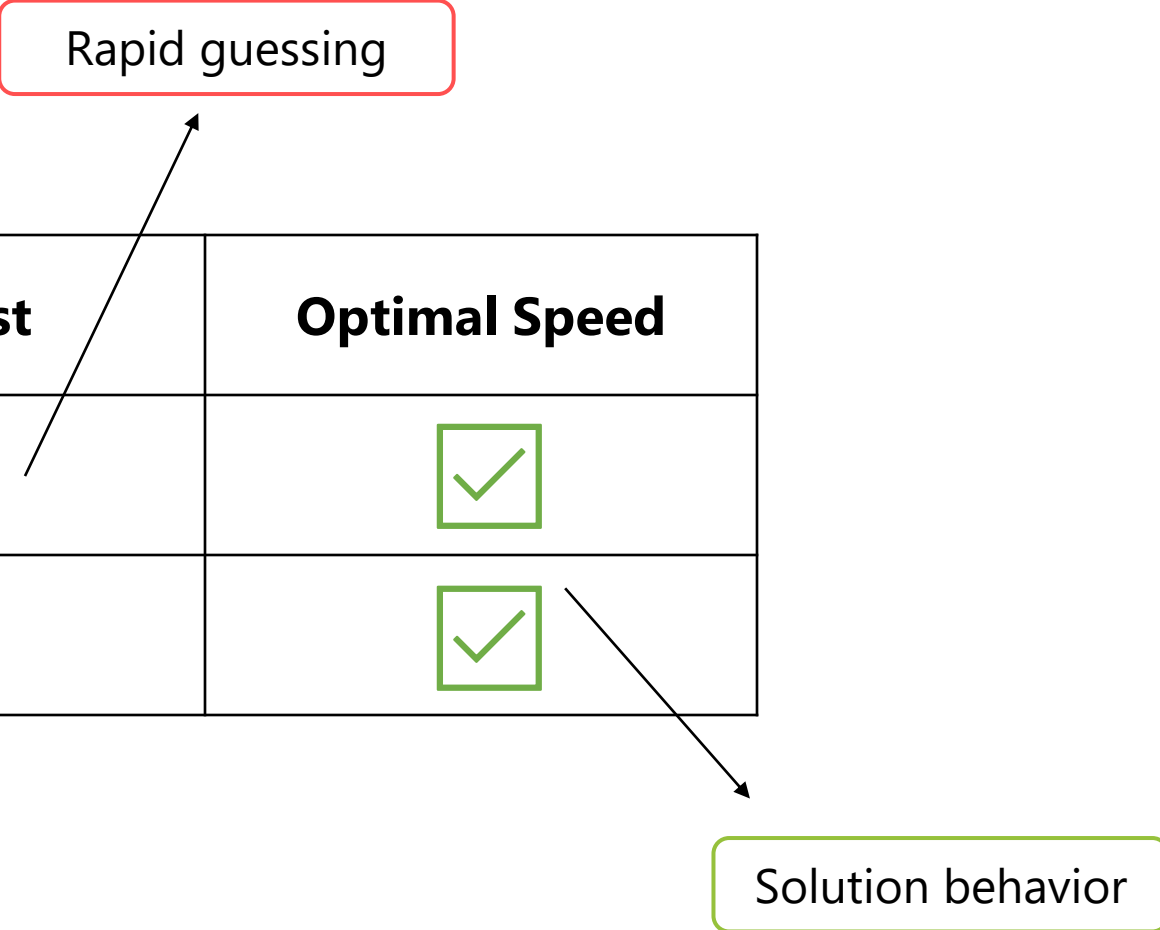


# Accuracy ~ Speed

	Too Fast	Optimal Speed
Correct		
Incorrect		

Rapid guessing

Solution behavior



# Operationalization of Solution Behavior

$$SB_{ij} = \begin{cases} 1 & \text{if } RT_{ij} \geq T_i \\ 0 & \text{otherwise} \end{cases}$$

where:

$SB_{ij}$  is the solution behavior indicator for person  $j$  on item  $i$ ,  $RT_{ij}$  is the response time of person  $j$  on item  $i$ , and  $T_i$  is a threshold for flagging rapid guessing for item  $i$ .

# Operationalization of Solution Behavior

$$P_{ij}(\theta) = (SB_{ij}) \frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}} + (1 - SB_{ij})(g_i)$$

where:

$g_i$  is the reciprocal of the number of response options for item  $i$   
(e.g.,  $1/4 = 0.25$  for a multiple-choice item with 4 options)

## **“High-Stakes” Assessments**

Rapid guessing → an attempt to maximize the score before the allotted time has expired

## **“Low-Stakes” Assessments**

Rapid guessing → a result of the lack of test-taking motivation

# **“Low-Stakes” Assessments**

Computerized formative assessments for K-12

Universal screening and progress monitoring

Early literacy, reading, and mathematics assessments

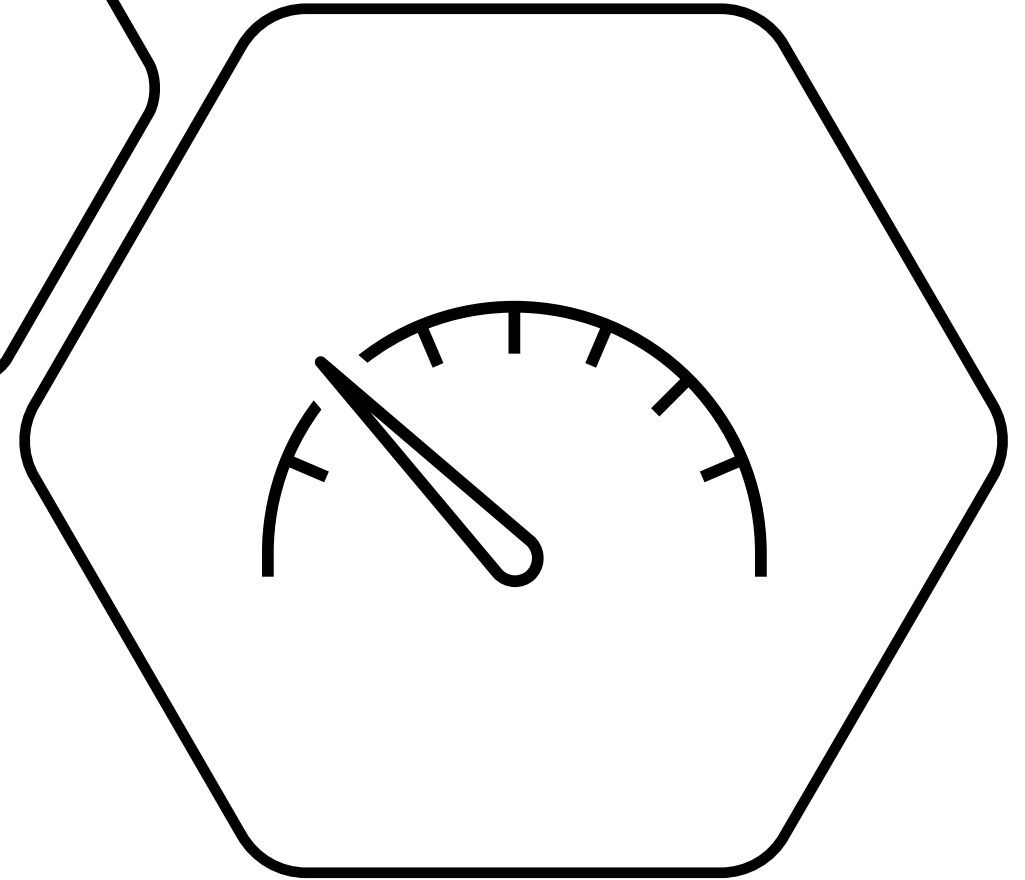
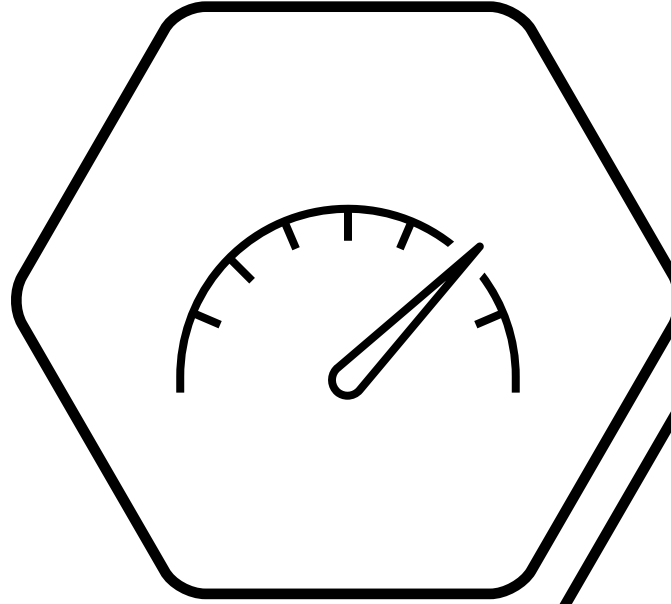
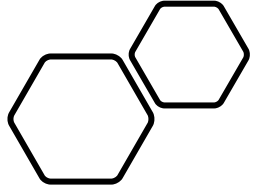
Seasonal (or weekly) linear or adaptive tests

Time limits (per item or test administration)

Automaticity (reading) and computational efficiency (math)

# Accuracy ~ Optimal Speed

	Too Fast	Optimal Speed	Too Slow
Correct			
Incorrect			



# Slow responding

Not being able to manage time on the test

- ☒ Failing to finish the test; not-reached items
- ☒ Rushing through the items in later positions
- ☒ Sacrificing ability to complete the test in time

# Solution Behavior: Revised

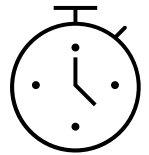
$$SB_{ij} = \begin{cases} 1 & \text{if } RT_{ij} \geq T_{i1} \text{ or } RT_{ij} \leq T_{i2} \\ 0 & \text{otherwise.} \end{cases}$$

where:

$SB_{ij}$  is the solution behavior indicator for person  $j$  on item  $i$ ,  $RT_{ij}$  is the response time of person  $j$  on item  $i$ , and  $T_{i1}$  and  $T_{i2}$  are the lower (rapid guessing) and higher (slow responding) thresholds for item  $i$ .



“Incorporating the speededness of the test into  
the operationalization of ability”  
(Tijmstra & Bolsinova, 2018, p.6)



Adopt a speed that allows you to complete the test



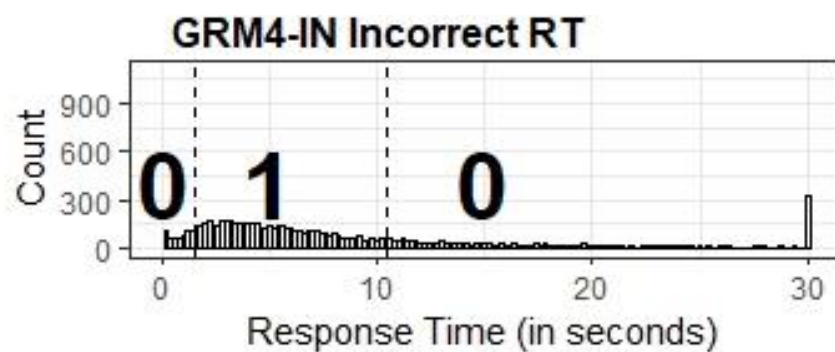
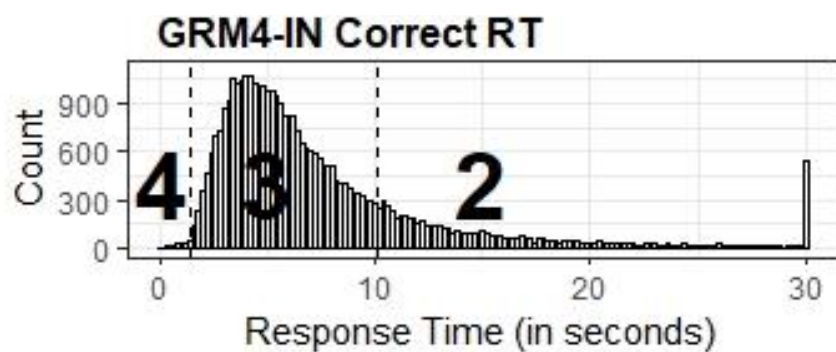
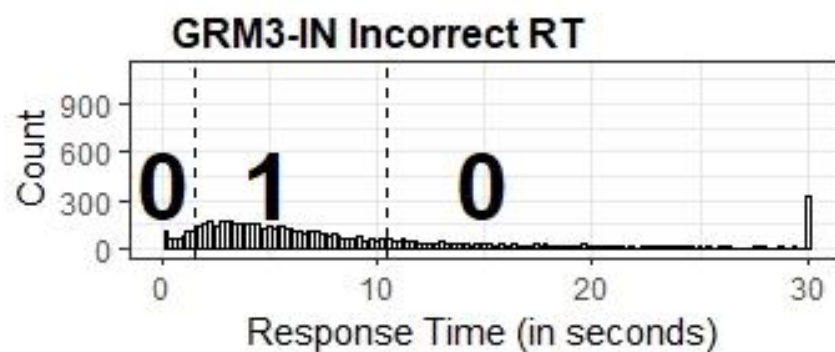
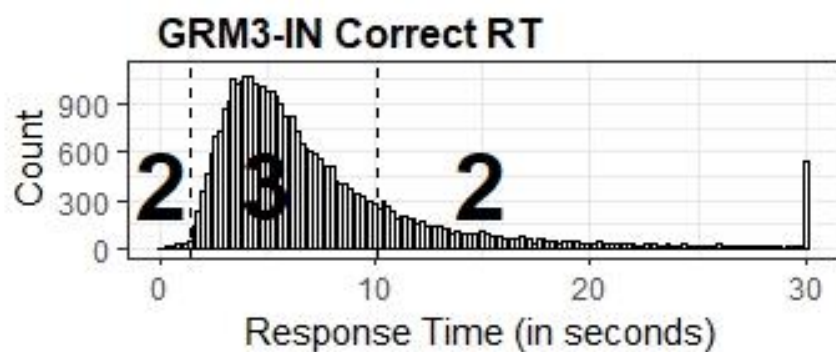
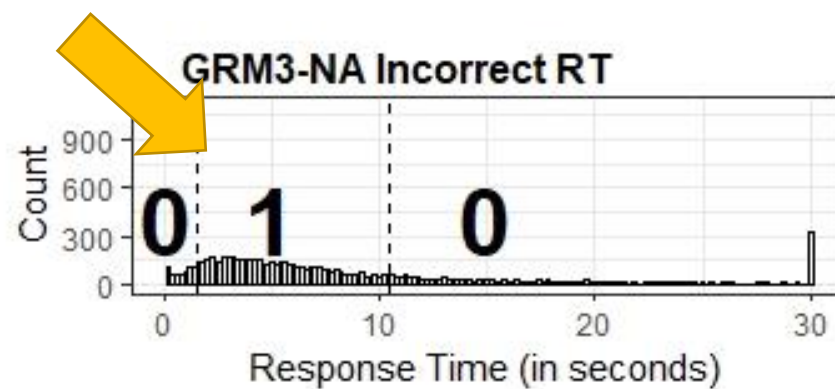
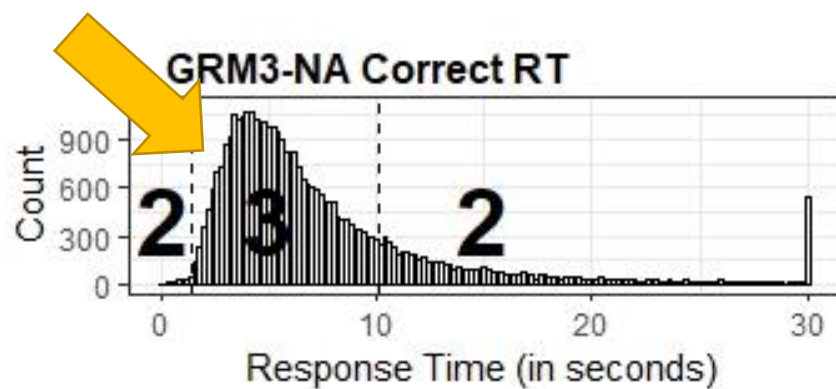
Perform to the best of your ability on the items

# Empirical Example (Gorgun & Bulut, 2021)

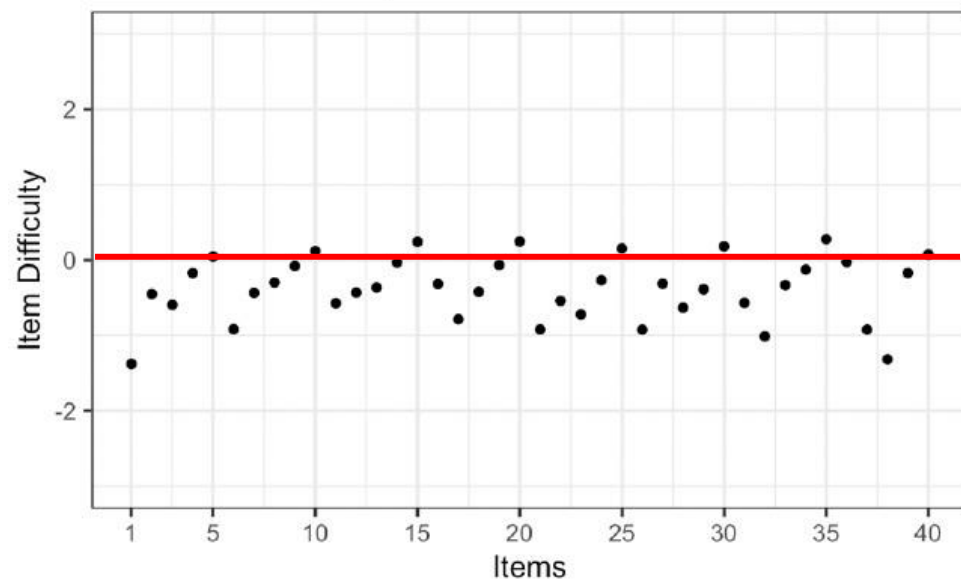
- 40,620 2<sup>nd</sup> and 3<sup>rd</sup> graders
- A computer-based, low-stakes math assessment
- Accuracy and fluency in basic math facts and operations (e.g., addition and subtraction)
- A maximum 30 seconds per item and 5-6 minutes for the entire test
- The percentage of not-reached items ranged from 11% to 68% across the items.

# Scoring and Item Calibration Procedures

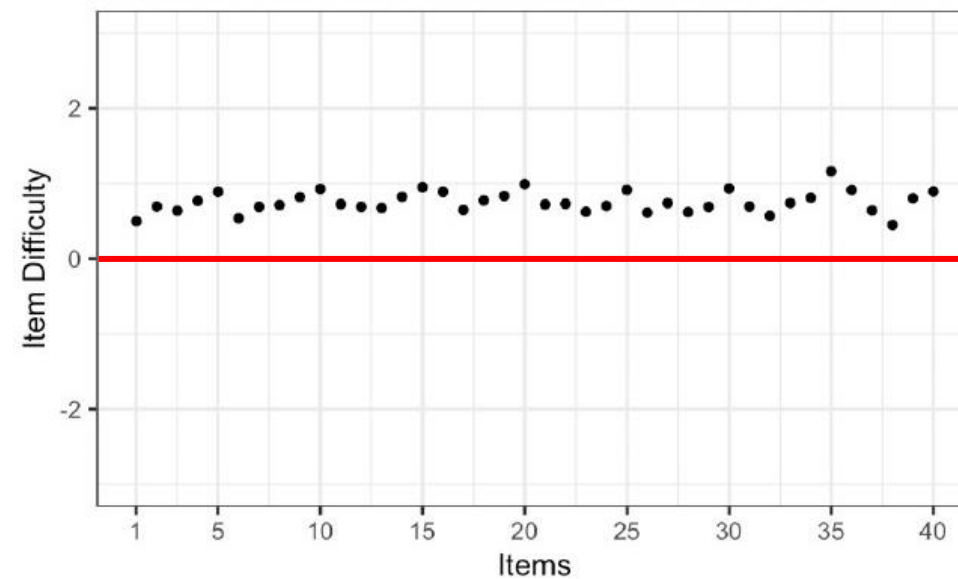
1. Recode not-reached as **NA** and apply the 2PL model (2PL-NA)
2. Recode not-reached as **0** and apply the 2PL model (2PL-IN)
3. Recode not-reached items as either **NA** or **0**, recode dichotomous item responses as **polytomous** based on optimal time use and, and apply the GRM
4. Response time thresholds using NT25 (Wise & Ma, 2012)
  - 25% and 175% of the median values of response time as normative thresholds
  - Separate thresholds for correct and incorrect responses



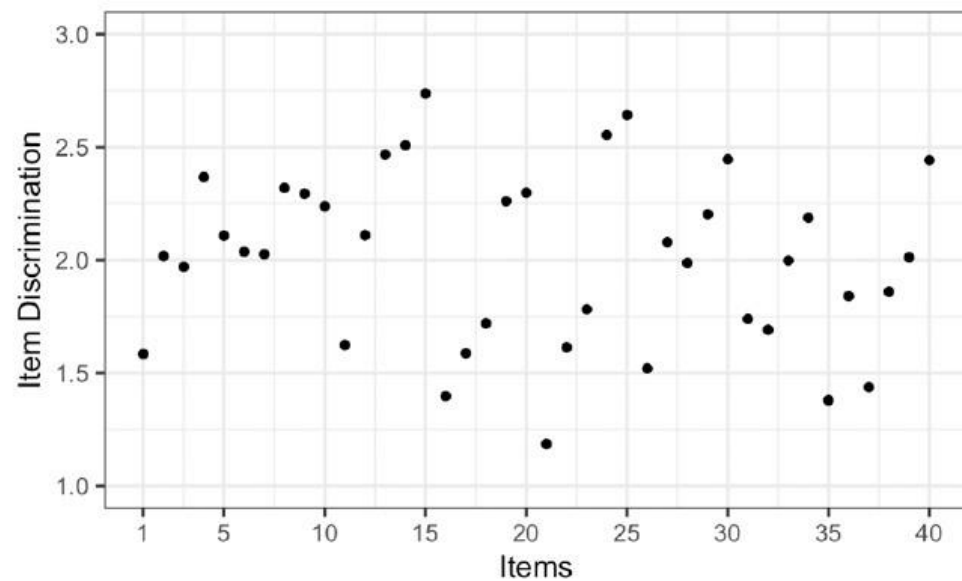
Estimated Difficulty Parameters for 2PL-NA



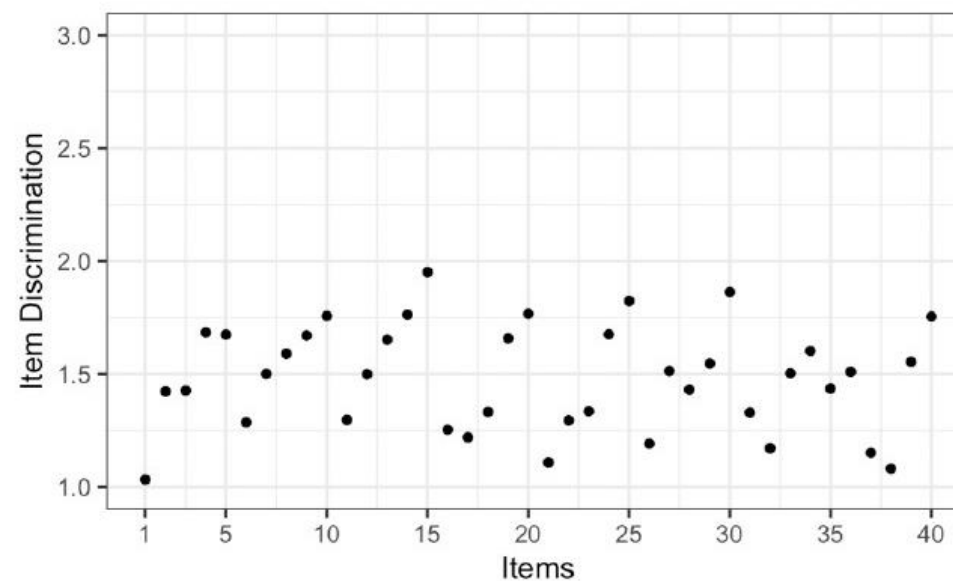
Estimated Difficulty Parameters for 2PL-IN



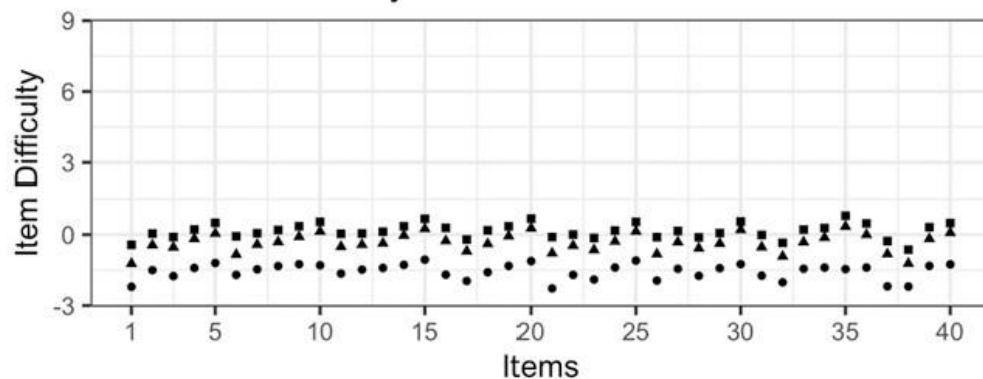
Estimated Discrimination Parameters for 2PL-NA



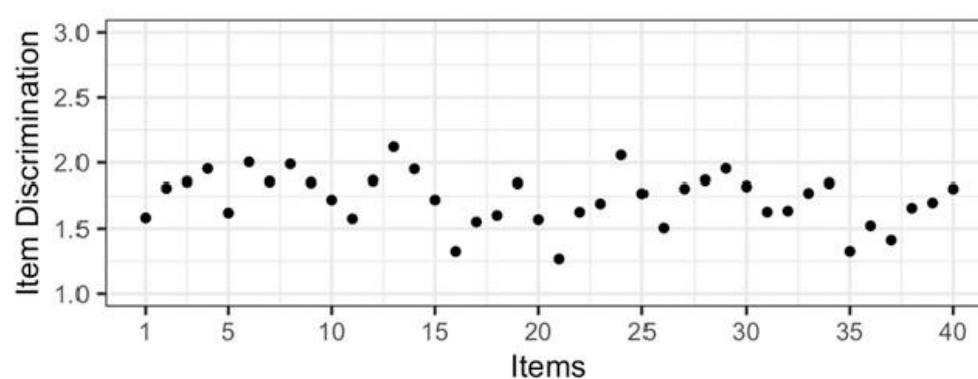
Estimated Discrimination Parameters for 2PL-IN



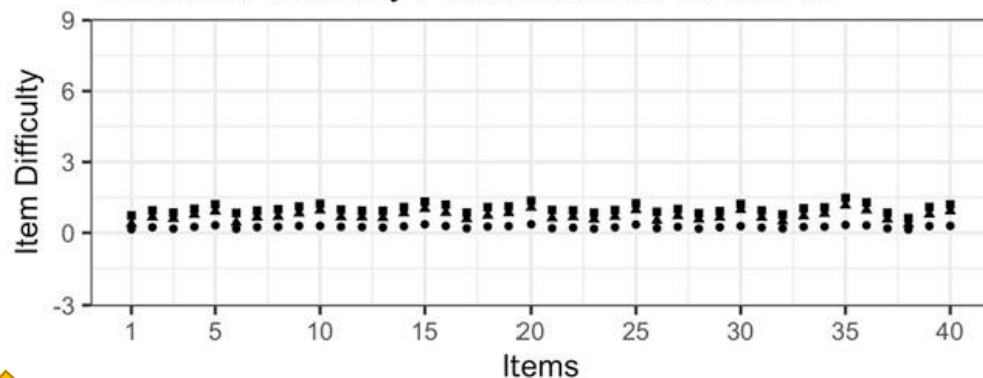
Estimated Difficulty Parameters for GRM3-NA



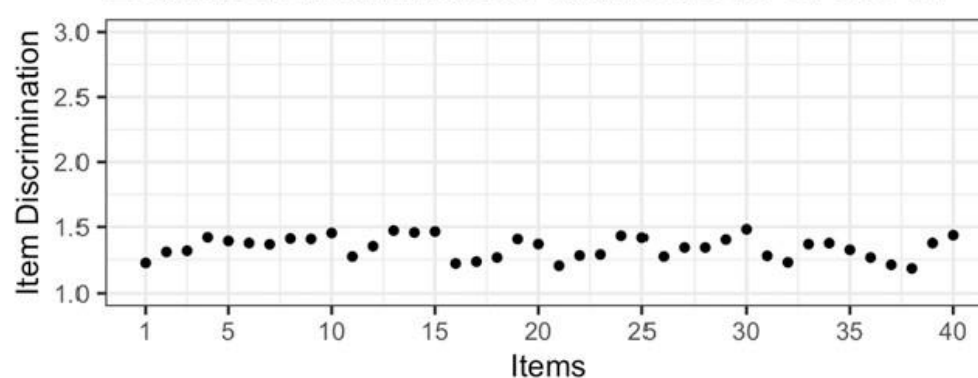
Estimated Discrimination Parameters for GRM3-NA



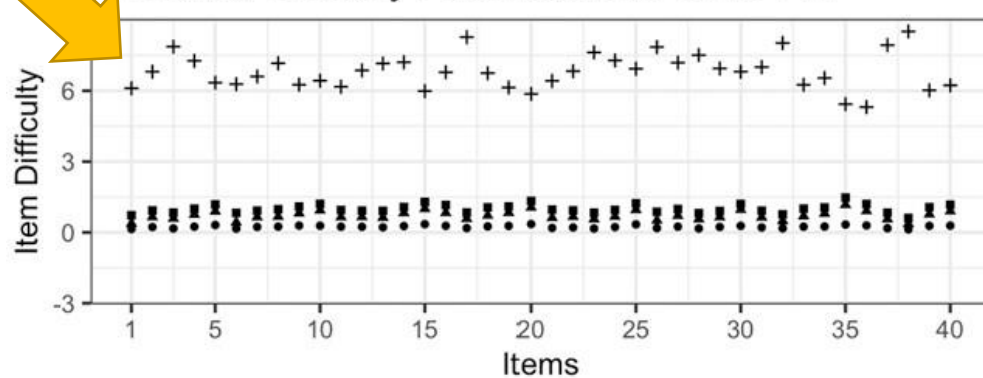
Estimated Difficulty Parameters for GRM3-IN



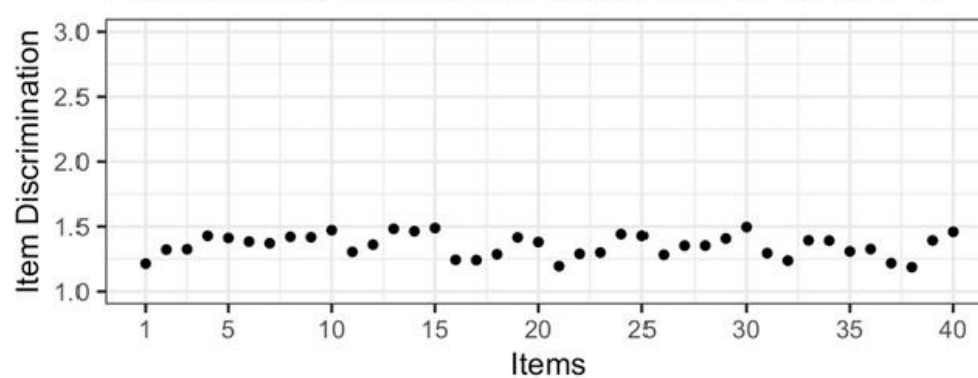
Estimated Discrimination Parameters for GRM3-IN

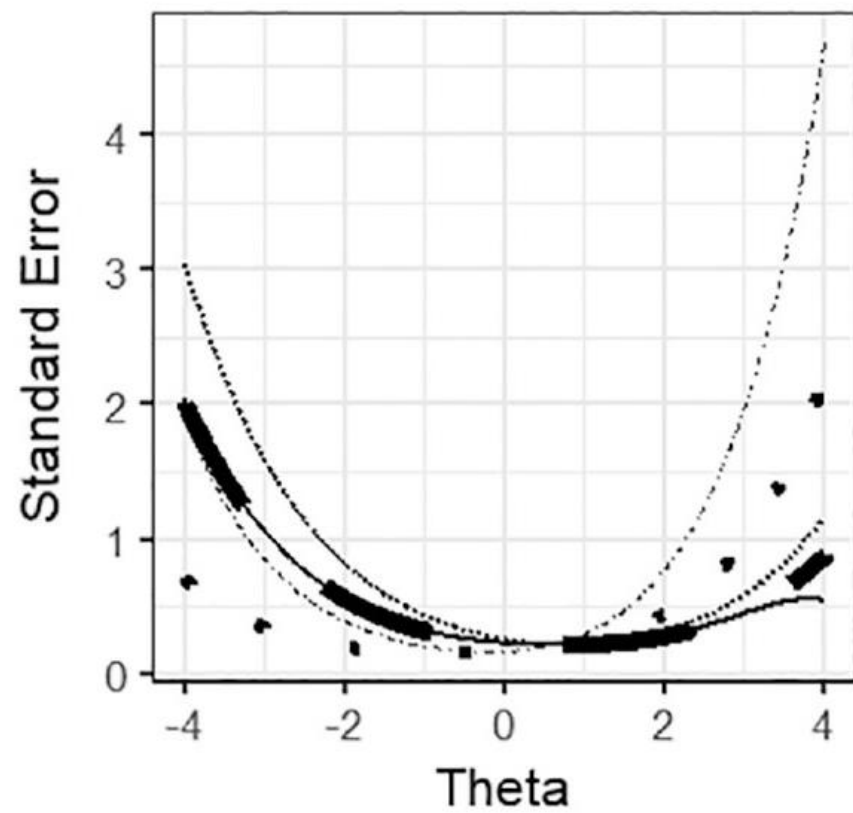
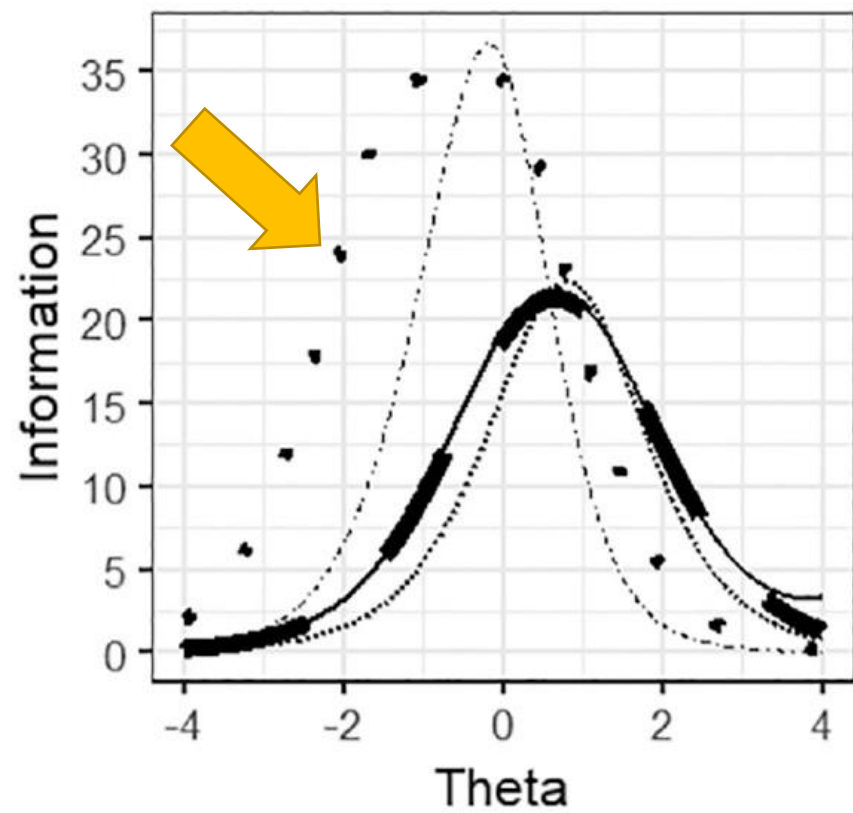


Estimated Difficulty Parameters for GRM4-IN



Estimated Discrimination Parameters for GRM4-IN





..... 2PL-IN    ---- 2PL-NA    — GRM3-IN    ■ GRM3-NA    — GRM4-IN



**Table 1.** Correlations Between the Estimated Ability Parameters From the Five Scoring Methods.

	2PL-NA	2PL-IN	GRM4-IN	GRM3-IN
2PL-IN	.858	—		
GRM4-IN	.571	.882	—	
GRM3-IN	.573	.883	.996	—
GRM3-NA	.903	.942	.822	.828

*Note.* 2PL = two-parameter logistic; GRM = graded-response model; 2PL-NA = 2PL model treating not-reached as missing; 2PL-IN = 2PL model treating not-reached as incorrect; GRM4-IN = GRM with four score points and not-reached treated as incorrect; GRM3-IN = GRM with three score points and not-reached treated as incorrect; GRM3-NA = GRM with three score points and not-reached treated as missing.

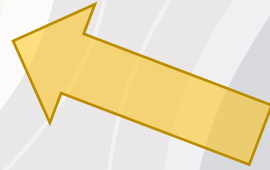


# Other Applications

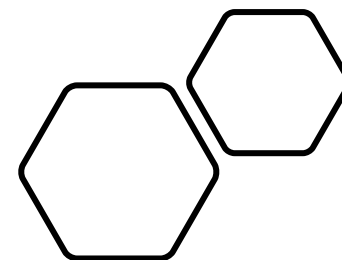
- Predicting test-taking disengagement in real-time using machine learning algorithms (Yildirim-Erbasli & Bulut, in press)
- Impact of test-taking disengagement on growth estimates from low-stakes educational assessments (Yildirim-Erbasli & Bulut, 2021)
- Using “speed” as a latent dimension when selecting items in computerized adaptive tests (Gorgun & Bulut, under review)



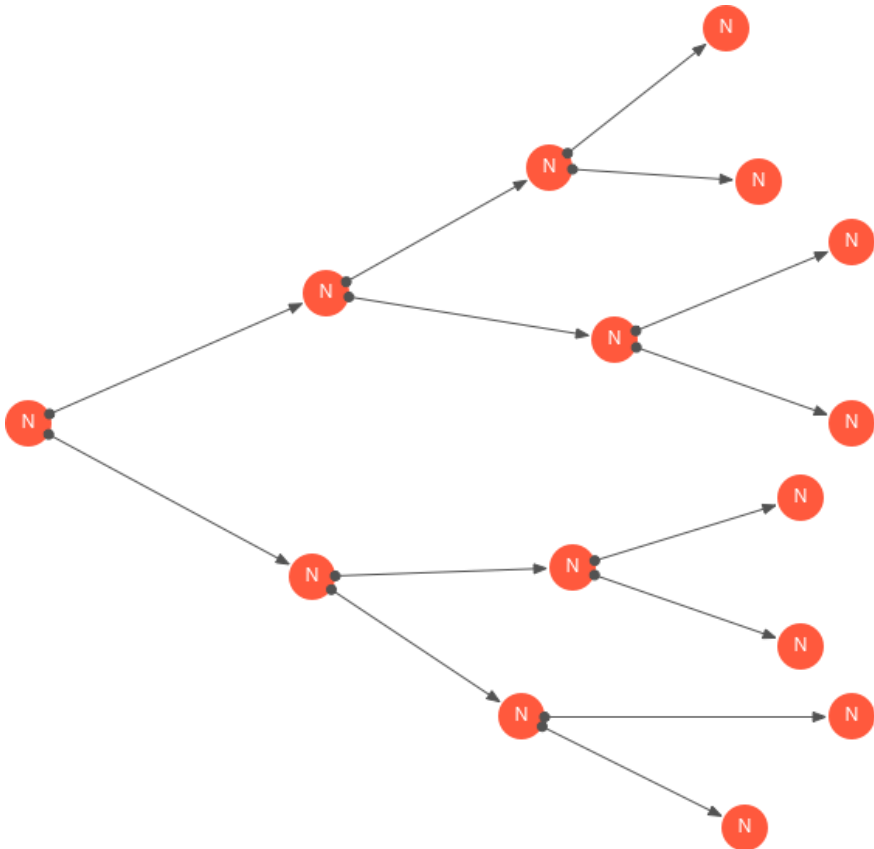
**Speed** as a latent  
dimension



**Speed** as a construct-  
irrelevant factor



# Adults' Web Navigation Strategies



Individuals may use different strategies when navigating through multiple web search results.

Multi-layered hypertext environments with top-layers (i.e., homepages) and deeper layers (i.e., nested pages)

Information foraging patterns

- Continue searching for information or terminate information acquisition process
- Trade-off between cognitive effort and efficient outcome



**Sampling**



**Breadth-First**

**Laborious**



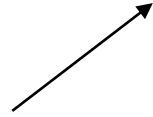
**Satisficing**



**Depth-First**

**Flimsy**

Satisficing & Depth-first

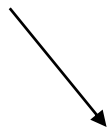


## Specific information-locating tasks

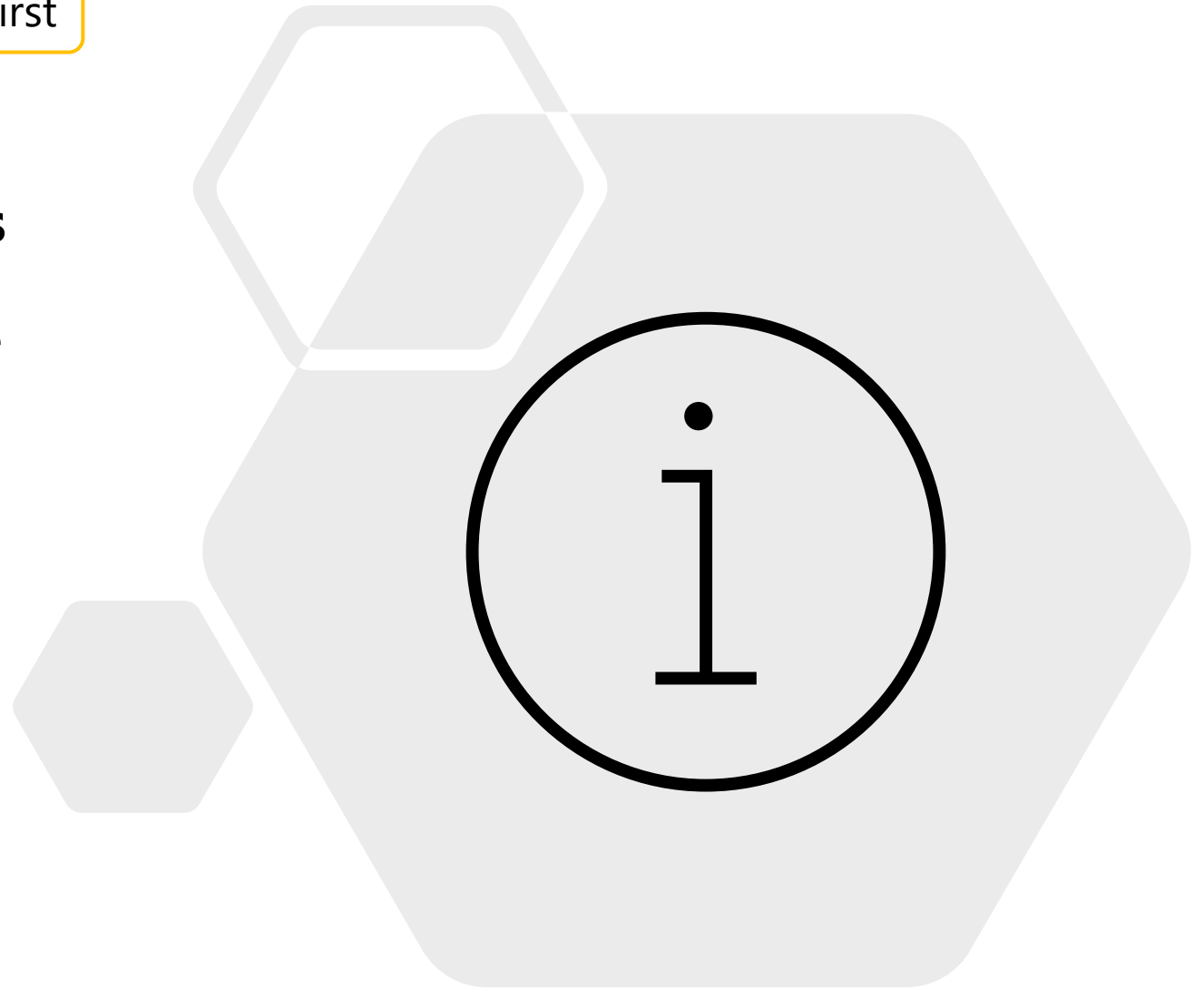
"Which online shop offers medications that could be used to treat the common cold in children and could be delivered in two days?"

## Information-evaluating tasks

"Which website do you think provides the most reliable treatment information for a common cold in children?"



Sampling







## Examining adults' web navigation patterns in multi-layered hypertext environments

Yizhu Gao<sup>a,\*</sup>, Ying Cui<sup>a</sup>, Okan Bulut<sup>a</sup>, Xiaoming Zhai<sup>b</sup>, Fu Chen<sup>c</sup>

<sup>a</sup> Department of Educational Psychology, University of Alberta, 6-110 Education Centre North, 11210 87, Ave NW, Edmonton, AB, T6G 2R5, Canada

<sup>b</sup> Department of Mathematics, Science, and Social Studies Education, University of Georgia, 110 Carlton Street, Athens, GA, 30602, USA

<sup>c</sup> Faculty of Education, University of Macau, Avenida da Universidade, Taipa, Macau, China

### ARTICLE INFO

#### Keywords:

Multi-layered hypertext environments  
Web navigation patterns  
Adult  
Clickstream data  
Latent class analysis  
Full-path sequence analysis

### ABSTRACT

Search engine users, when presented with multiple web search results, must be able to flexibly navigate dependent on the type of search tasks. Users who demonstrate inappropriate navigation behavior would likely fail to obtain target information. However, very few studies have examined users' web navigation patterns and their effectiveness in completing different types of search tasks in multi-layered hypertext environments. To fill these research gaps, we used data composing a sample of 1408 adults in the United States and the United Kingdom who completed both a specific information-locating task and an amorphous information-evaluating task when participating in the Programme for the International Assessment of Adult Competencies (PIAAC) 2012. We extracted participants' clickstream data as well as their performance on the two tasks from recorded web navigation logs. Results from a series of latent class analyses and full-path sequence analyses showed five distinct behavioral patterns when participants performed the two tasks: of which, *Flimsy*, *Breadth-first*, *Laborious*, and *Sampling* patterns were revealed in both tasks, while the *Satisficing* pattern emerged exclusively in the specific information-locating task. Regarding pattern effectiveness, the *Sampling* pattern group outperformed other navigation pattern groups in the amorphous information-evaluating task, whereas the *Satisficing* pattern group performed best in the specific information-locating task.

### 1. Introduction

In today's global information society, searching for information on the World Wide Web (WWW) to solve real-world problems has increasingly evolved into an indispensable part of people's everyday life. [Internet World Stats \(2021\)](#) estimates that 65.6% of the total population around the world are Internet users and a large proportion of these people frequently utilize the WWW for information-seeking purposes. For example, people who have medical problems may use Internet search engines to initiate search queries for treatments. After entering search terms in a selected search engine (e.g., Google) and sending them off, people are usually exposed to a so-called search engine results page (SERP) with a list of search result links. Diverse search results contain a mixture of information with variant specificities such as timeliness and authority, which makes online information at different levels of usefulness, suitability, and trustworthiness to information problems. In consideration of their information needs such as acquiring the most

professional and/or the most up-to-date medical advice, web users navigate search results to identify target information ([Brand-Gruwel et al., 2009](#); [Walraven et al., 2009](#)).

An important distinction can be made between single- and multi-layered hypertext environments in which web information is distributed ([Chang & Chen, 2011](#); [Cho & Afllerbach, 2017](#)). In the case of single-layered hypertext environments, information is visibly displayed on homepages after web users click on search results links on SERPs. In contrast, navigating for desired information in multi-layered hypertext environments is more complex because the information is typically distributed through multi-layered hypertexts. This means that some information, which might be both helpful and necessary to solve information problems, is hidden beneath a series of hypertexts. As a result, web users engage in a multi-layered navigation process to access information that is contained in top-layered homepages and deeper-layered, nested web pages ([Coiro & Dobler, 2007](#)). It should be noted that a multi-layered structure is the primary feature of web

\* Corresponding author.

E-mail addresses: [yizhu@ualberta.ca](mailto:yizhu@ualberta.ca) (Y. Gao), [yc@ualberta.ca](mailto:yc@ualberta.ca) (Y. Cui), [bulut@ualberta.ca](mailto:bulut@ualberta.ca) (O. Bulut), [Xiaoming.Zhai@uga.edu](mailto:Xiaoming.Zhai@uga.edu) (X. Zhai), [fu4@ualberta.ca](mailto:fu4@ualberta.ca) (F. Chen).

<https://doi.org/10.1016/j.chb.2021.107142>

Received 2 April 2021; Received in revised form 3 December 2021; Accepted 8 December 2021

Available online 11 December 2021

0747-5632/© 2021 Elsevier Ltd. All rights reserved.



# PIAAC

Programme for the International  
Assessment of Adult Competencies

<https://www.oecd.org/skills/piaac/>

81,744 adults between the ages of 16  
and 65 from 17 countries

Problem-solving in technology-rich  
environments (PSTRE) via 14  
computerized tasks

The US ( $n = 659$ ) and the United  
Kingdom ( $n = 1182$ ) samples were used.

## 1. The book purchase task (specific-information-locating task)

- **Task:** Finding the best book choice for a friend's birthday using six websites.
- **Criteria:** 1) The shipping date should be within two weeks, 2) the target audience should be beginners of photography, and 3) the price should be less than \$40.

## 2. The reliable site task (information-evaluating task)

- **Task:** Recommending one of the five websites to a friend as the most reliable and trustworthy treatment information for a sprained ankle
- **Criteria:** Not specified (the authority of information providers and purposes of shared information)



Unit 1 - Part 1

You are looking for a job and have located these five websites.

You want to use a site that does not require you to register or pay a fee.

Bookmark all the sites that meet your requirements.

Once you have bookmarked the sites, click Next to go on.



Web
File Edit Bookmark Help

URL:

Web Search

[Find Your Job - JobSearch.com](#)  
The best job search site on the web. Check with us first!  
[www.jobsearch.com](http://www.jobsearch.com)

[Job Links](#)  
We connect you with the best jobs on the web.  
[www.joblinks.com](http://www.joblinks.com)

[Looking for a job?](#)  
Start your job search here.  
[www.careerstarters.com](http://www.careerstarters.com)

[Connections.com](#)  
We provide access to the best jobs  
[www.connections.com](http://www.connections.com)

[The best jobs online](#)  
If you are looking for the perfect job, start right here.  
[www.greatjobs.com](http://www.greatjobs.com)

Web

The links structure and the corresponding web pages for the two tasks.

Task	SERP					
	Link 1	Link 2	Link 3	Link 4	Link 5	Link 6
Book Purchase	H1 N1	H2	H3 N3	H4 N4	H5 N5	H6 N6
Reliable Site	H1 N1	H2 N2	H3 N3	H4	H5	

*Note.* H indicates the homepage. N denotes the nested web page.

# Analysis Framework

## **Latent class analysis**

Identify groups that displayed different navigation patterns

## **Full-path sequence analysis (Gabadinho, Ritschard, Müller, & Studer, 2011)**

Identify typical behavioral patterns that best characterize the set of sequences of each navigation pattern group (using the [TraMineR](#) package in R)

# Dissimilarities Between Sequences (Optimal Matching)

$$d(s_1, s_2) = A(s_1, s_1) + A(s_2, s_2) - 2A(s_1, s_2)$$

where:

- $d(s_1, s_2)$  is the distance between sequences of  $s_1$  and  $s_2$
- $A(s_1, s_2)$  is the count of common attributes between  $s_1$  and  $s_2$

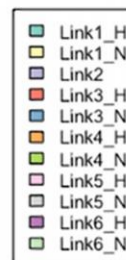
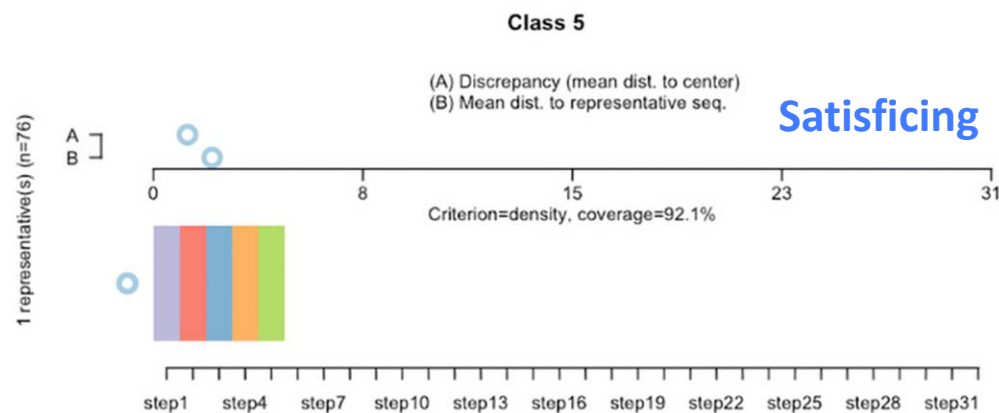
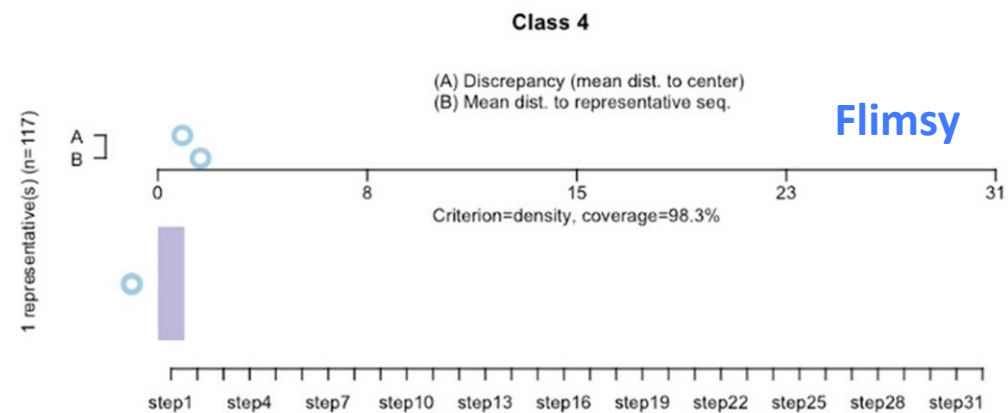
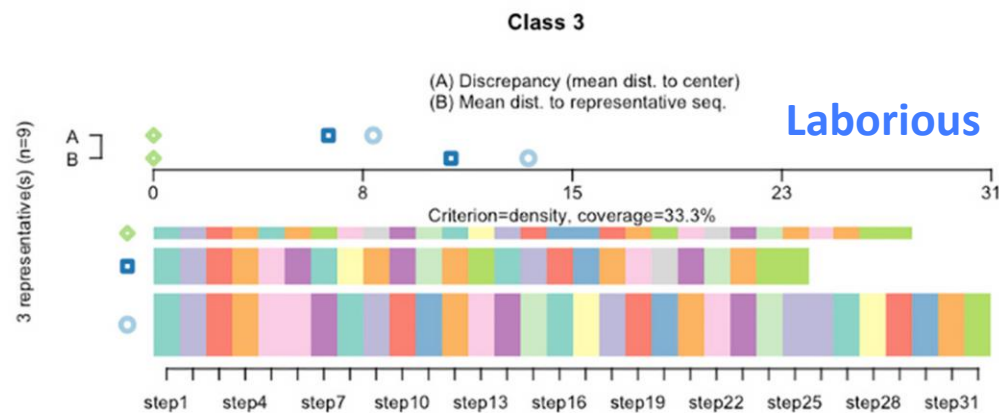
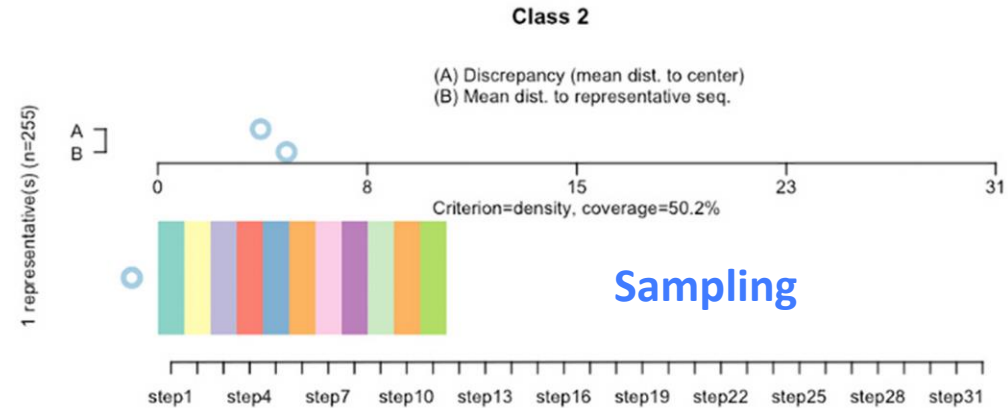
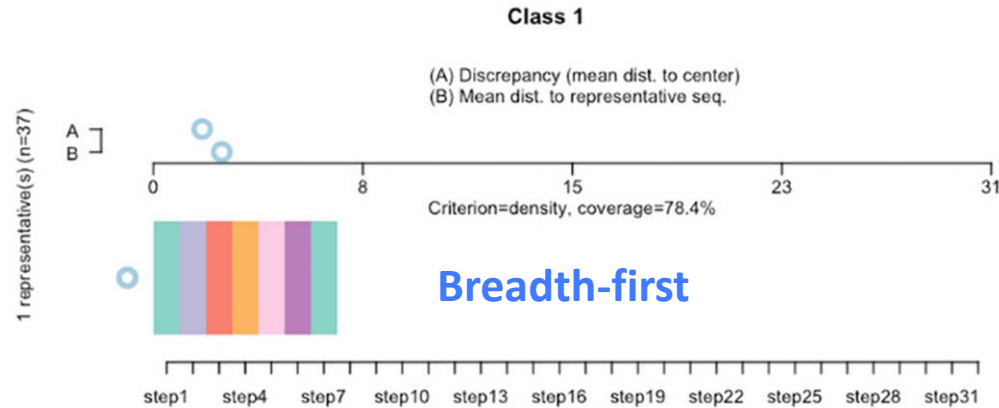
$d(s_1, s_2)$  gets smaller as two sequences become more similar.

1. Calculate a pairwise distance matrix with all  $d(s_1, s_2)$  values
2. Count by row (or column) the number of distances that are less than a defined threshold (called *neighbourhood density*)
3. Sort all distinct sequences according to their neighbourhood density
4. Eliminate redundancy based on a proportion of the maximal theoretical distance ( $D_{max}$ ):

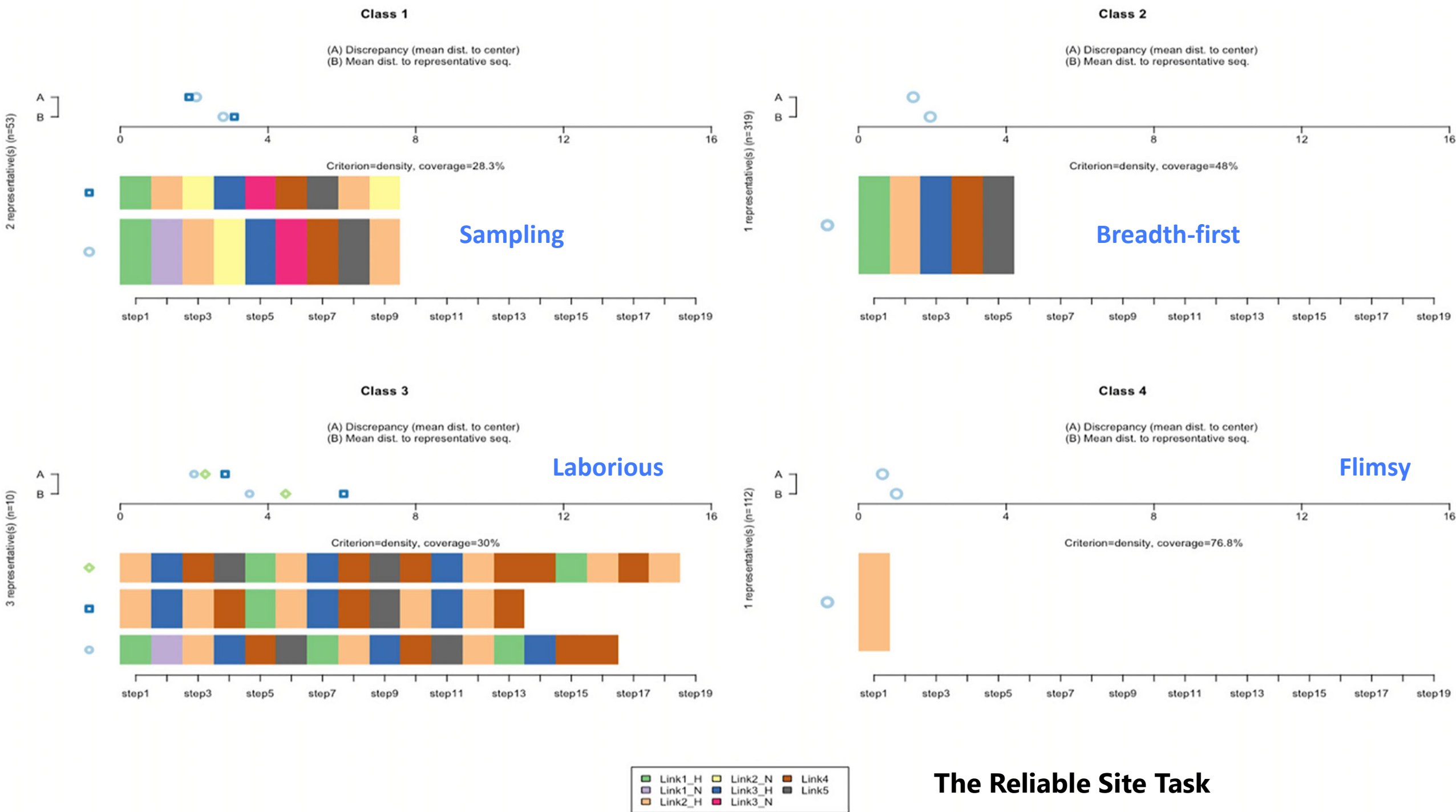
$$D_{max} = \min(\ell_1, \ell_2) \cdot \min(2C_I, \max(S)) + |\ell_1 - \ell_2| \cdot C_I$$

where:

$\ell_1$  and  $\ell_2$  are the lengths of two sequences  $s_1$  and  $s_2$ ,  $C_I$  is the indel cost, and  $\max(S)$  is the maximal substitution cost.



**The Book Purchase Task**



Sample size and correctness of navigation patterns for two tasks.

Navigation Pattern	United States				United Kingdom			
	Book Purchase Task		Reliable Site Task		Book Purchase Task		Reliable Site Task	
	Size	Correctness (%)	Size	Correctness (%)	Size	Correctness (%)	Size	Correctness (%)
Flimsy	117	0	112	37.50	215	14.42	171	38.60
Breadth-first	37	0	53	42.95	96	0	630	56.80
Laborious	9	66.67	10	60.00	23	78.26	NA	NA
Sampling	255	84.71	319	92.45	487	81.93	113	92.90
Satisficing	76	96.05	NA	NA	93	93.55	NA	NA

Note. NA: Not applicable.



# Future Directions

- The issue of preprocessing “process data”
  - Theory- and data-driven reasoning from log data (Goldhammer et al, 2021)
  - A framework for purposeful data collection (Mostow & Beck, 2009)
- Speed for more “personalized” assessments
  - Speed of typing answers for constructed-response items
- Using [transformer](#) models (e.g., BERT) to analyze sequential log data
  - [cyBERT](#)
  - [logBERT](#)



# Thank You!

For questions/comments:

[bulut@ualberta.ca](mailto:bulut@ualberta.ca)