GlassPy Process Data Analytics Solution and its Applications at Educational Testing Service

Jiangang Hao
Educational Testing Service
6/18/2020

About ETS and Process Data

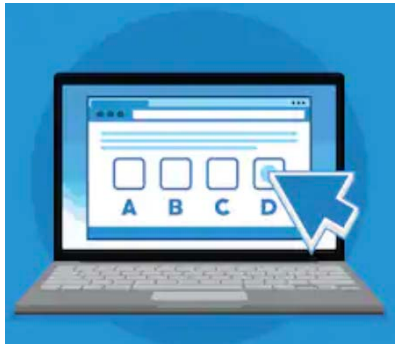
- ETS, funded in 1947, is a large non-profit testing company headquartered in Princeton, New Jersey, USA.
- We carry out about 50 million tests worldwide every year. Some well- known tests includes GRE, TOEFL, and SAT.
- We have a large R&D division, doing research on psychometrics, statistics, natural language processing, reading, writing, etc.
- https://www.ets.org/research

- ETS scientists work on process data in various programs and projects:
  - PISA and PIAAC
  - NAEP
  - Winsight
  - Writing analytics
  - Collaborative problem solving
  - Remote testing
  - …
- Assessment Analytics and Data Science: a group dedicated to research on process data

Process Data and their Uses

Process data capture fine-grained interaction information between learners and learning/assessment tasks, allowing better understanding of the learners and the tasks

People



Tasks

Typical uses

- Improve psychometric quality
- Evidence for process-based constructs
- Group difference and fairness
- Test security
- Feedback to learners and other stakeholders
- Improve the tasks and delivery platform

Challenges of Process Data in Practice

- Engineering aspect
  - Capturing, hosting, transferring, and manipulating data with proper software infrastructure, tools and techniques.
  - Data engineer/Architect
- Scientific aspect
  - Making sense of data to realize their potential value for intended interpretations and uses
  - Data scientist, Psychometrician, Statistician
- Organizational aspect
  - Workforce and professional development
  - Identifying suitable workflow (e.g., agile, waterfall, etc.)

GlassPy Analytics Solution at ETS

- GlassPy: Game Log Analysis in Python, started in 2014 at ETS
- We envision glassPy more as a generic framework for handling process data from digital-based assessments (DBAs)
- It is an "adaptor" to connect the Evidence Centered Design (ECD) to data in practice
- GlassPy analytics solution Includes three main components

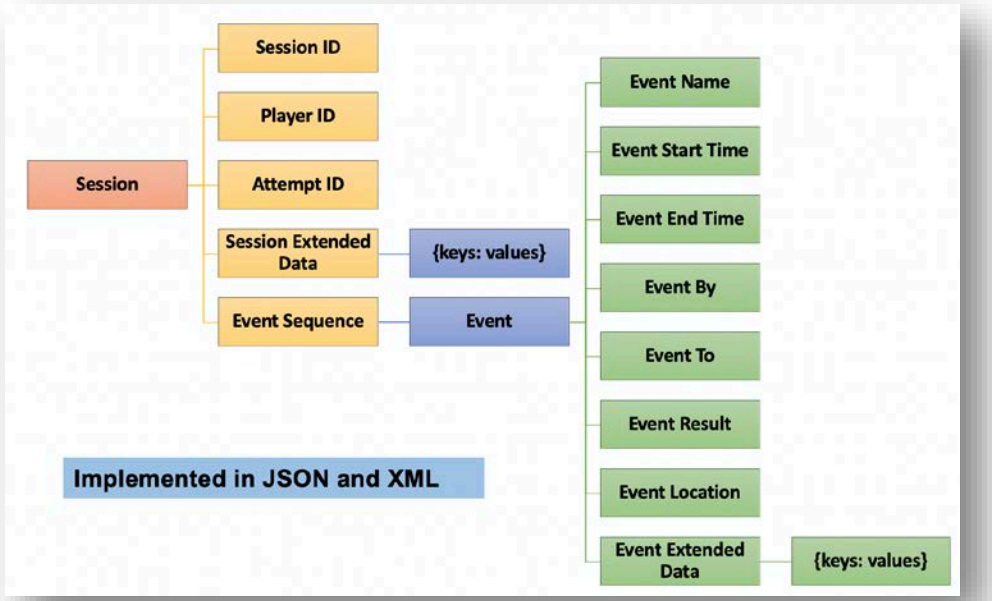
Evidence Identification Centered Data Design - EICDD

GlassPy Library

Interactive Analytics Frontend

Evidence Identification Centered Data Design

Data Model



Standard Operational Procedure

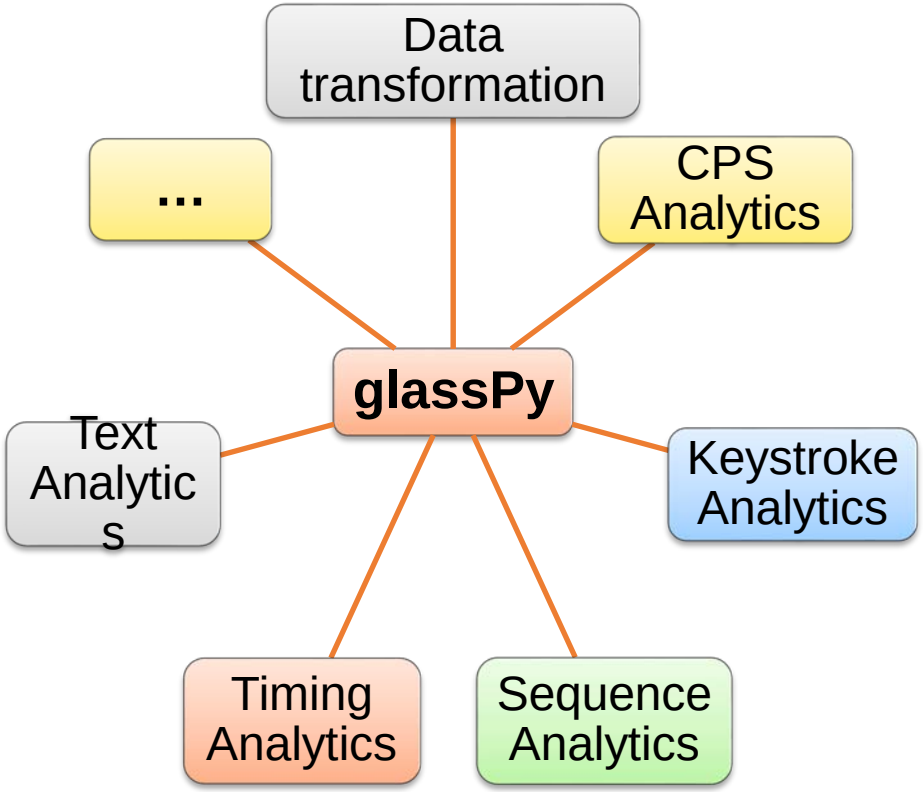
X: participant XX: coordinator

		Roles				
		Learning Scientist/ Assessment Developer	Psychometrician	Programmer	Data Scientist	Data analyst
Timeline	**Step 1**: All-party meeting to brief the procedures	X	X	X	XX	X
	Step 2: Specification of "Q matrix" for constructs and evidences	X				
	Step 3: Specification of additional evidence for psychometric modeling need		X			
	Step 4: Translate the evidences into operable task codes			X	X	XX
	Step 5: Design log file structure and schema				X	
	Step 6: Implement the evidence logging into the log file and tryout			X	X	XX
	Step 7: Log file parsing and evidence extraction based on the tryout data				X	
	Step 8: QA the log files to check whether the recovered evidence is sufficient. If not, restart from **Step 6**. If everything is good, proceed to next step	X	X		X	XX
	Step 9: Finalize the data reduction pipeline and QA procedure				XX	X
	Step 10: Implement the data reduction				X	XX

Evidence Trace File

Hao et al., 2016; Hao & Mislevy, 2018

GlassPy Library

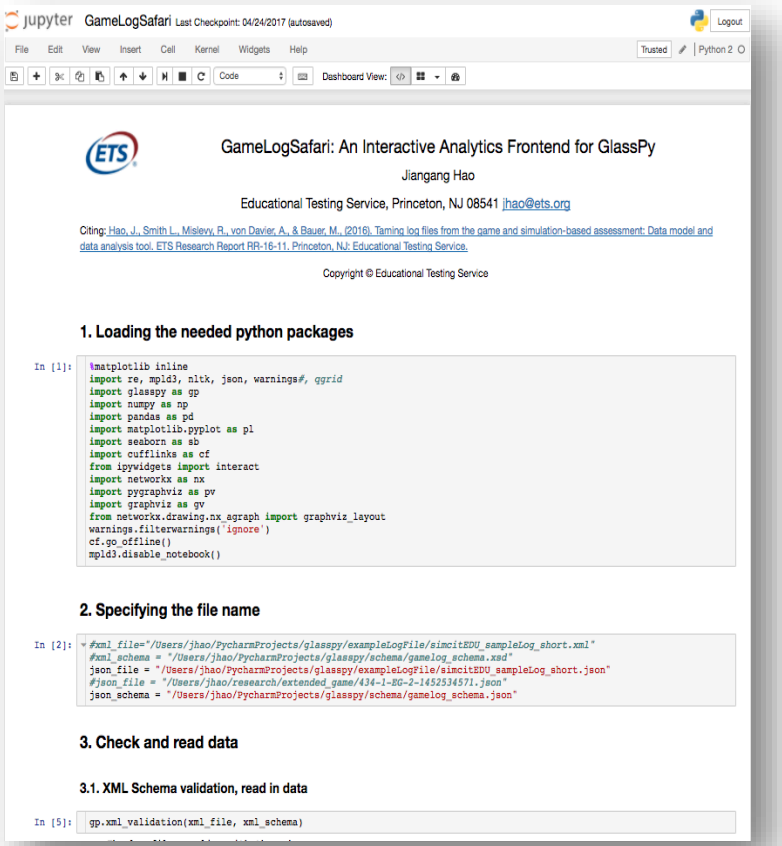


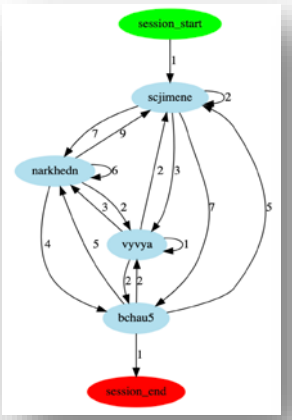

```
def find_sub_str(sl, l):
    """
    Purpose: return the location of the sub-string/list in a string/list for multiple matching::

        Syntax: find_sub_str(sl, l)

        Input: sl sub-string/list, l: string/list

        Return: location of the sub-string/list in a string/list.

        Created By: J. Hao @ 2014
    """
    sl = list(sl)
    l = list(l)
    results = []
    sll = len(sl)
    for ind in (i for i, e in enumerate(l) if e == sl[0]):
        if l[ind:ind + sll] == sl:
            results.append([ind, ind + sll - 1])
    return results


def ngram_freq(event_list, n=1, sorted=False):
    """
    Purpose: calculate the frequency of ngrams of the event list ::

        Syntax: ngram_freq(eventList,ngram=1)

        Input: GameDataFrame

        Output: Dictionary of ngram and the frequencey

        Created By: J. Hao @ 2014
    """
    if event_list is None:
        return 'Syntax: ngramFreq(eventList, ngram)'
    else:
        event_list = list(event_list)
        ngm = ngrams(event_list, n)
        ngm = cl.Counter(ngm)
        df = pd.DataFrame([list(ngm.keys()), list(ngm.values())])
        df = df.T
        df.columns = ['events', 'freq']
        if sorted:
            return df.sort_values(by='freq', ascending=False)
        else:
            return df


def regularize_htmltext(res):
    reg_response = re.sub('<[^<]+?>', ' ',res).replace(' ',' ').replace('\xc2\xa0',' ')
    reg_response = re.sub('[\x80-\xff]',' ',reg_response)
    return reg_response
```

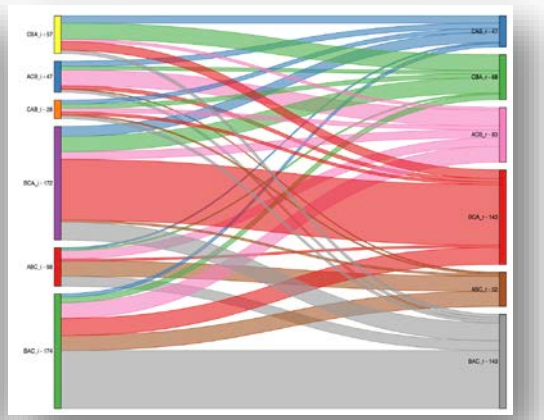
Not open sourced yet and under active expanding

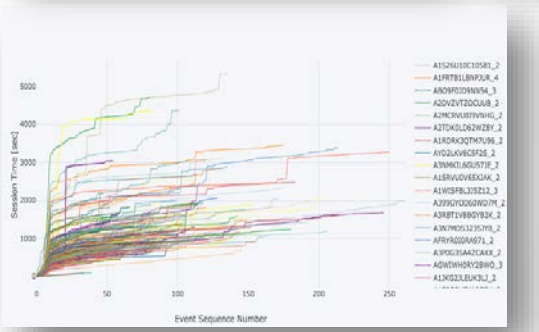
Interactive Analytics Frontend

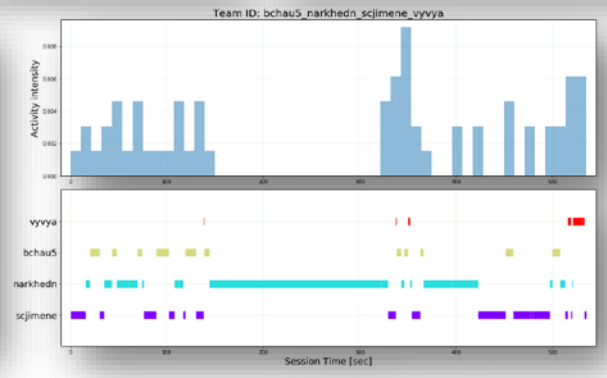
Jupyter/Zeppelin Notebook/Lab + Plotly Dash + Shiny + FastAPI + others







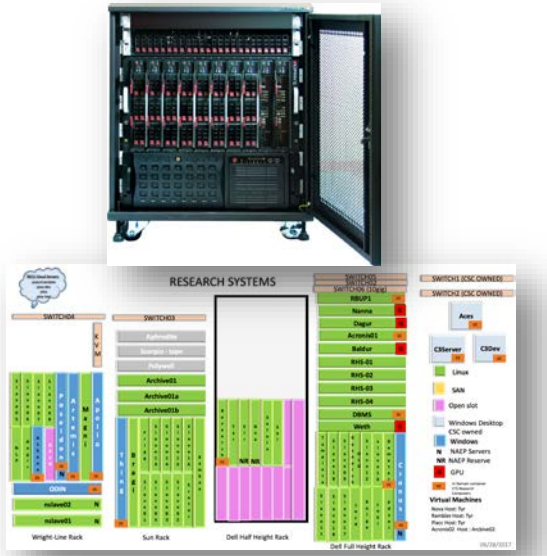




Scaling up in Operational Work

Individual computers

Computing grids

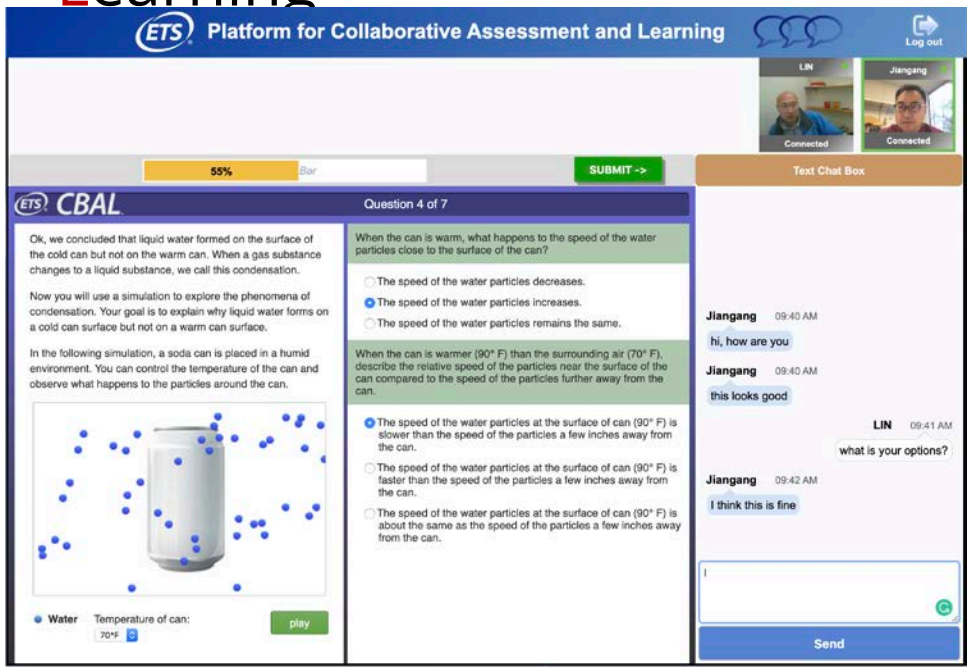


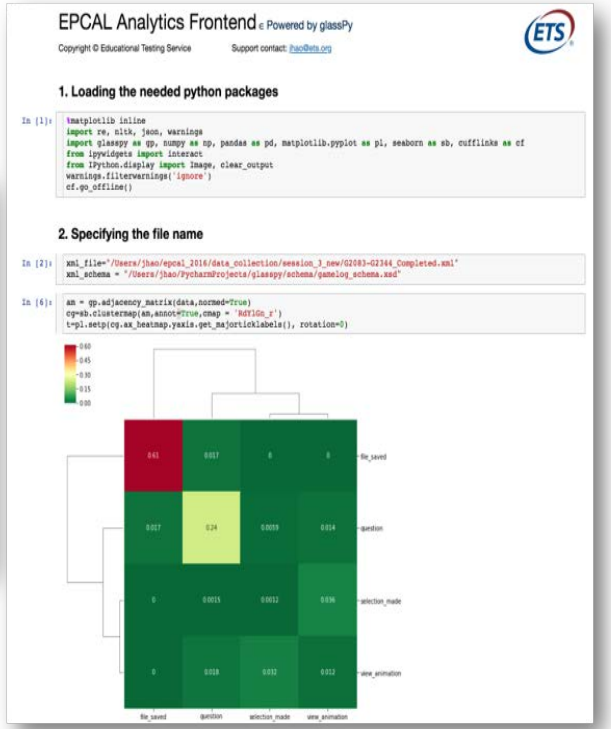
Cloud computing

git

A Full Implementation in EPCAL

Fully implemented in the ETS Platform for Collaborative Assessment and Learning





Analytics vs. Modeling

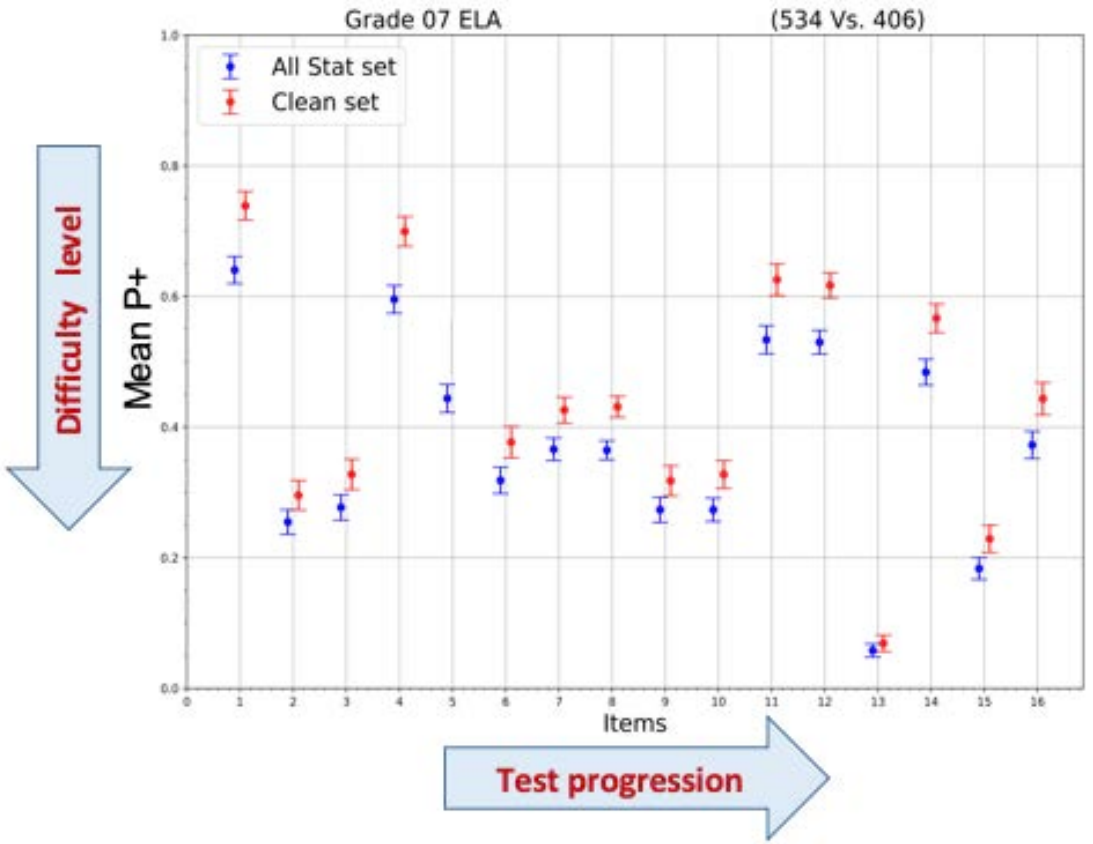
- Analytics
  - Discovery, interpretation, and communication of meaningful patterns in data and apply these patterns towards effective decision making.
  - Suited for data from most digital learning and assessment tasks

- Modeling
  - Psychometric/statistical modeling (e.g., Bayes net, IRTs, CDM, many others)
  - Cognitive modeling (e.g., ACT-R, SOAR, many others)
  - Better suited for data from carefully designed virtual performance tasks

- Complication in real-world data
  - Response process is highly regulated by task/item design, limiting the generalizability of methods.
  - Stationary condition generally does not hold throughout, requiring regime switching.
  - The response sequences span a big and sparse mathematical space, making it challenging to identify suitable statistical distributions.
  - The response process are often short, challenging the parameter estimation.

Hao & Mislevy, (2020), A data science perspective of next generation psychometrics

Example1: process data for item calibration



Definitions:

- Clean set: students who complete the assessment in a single session
- All Stat set: students who complete the assessment in a single session and in multiple sessions

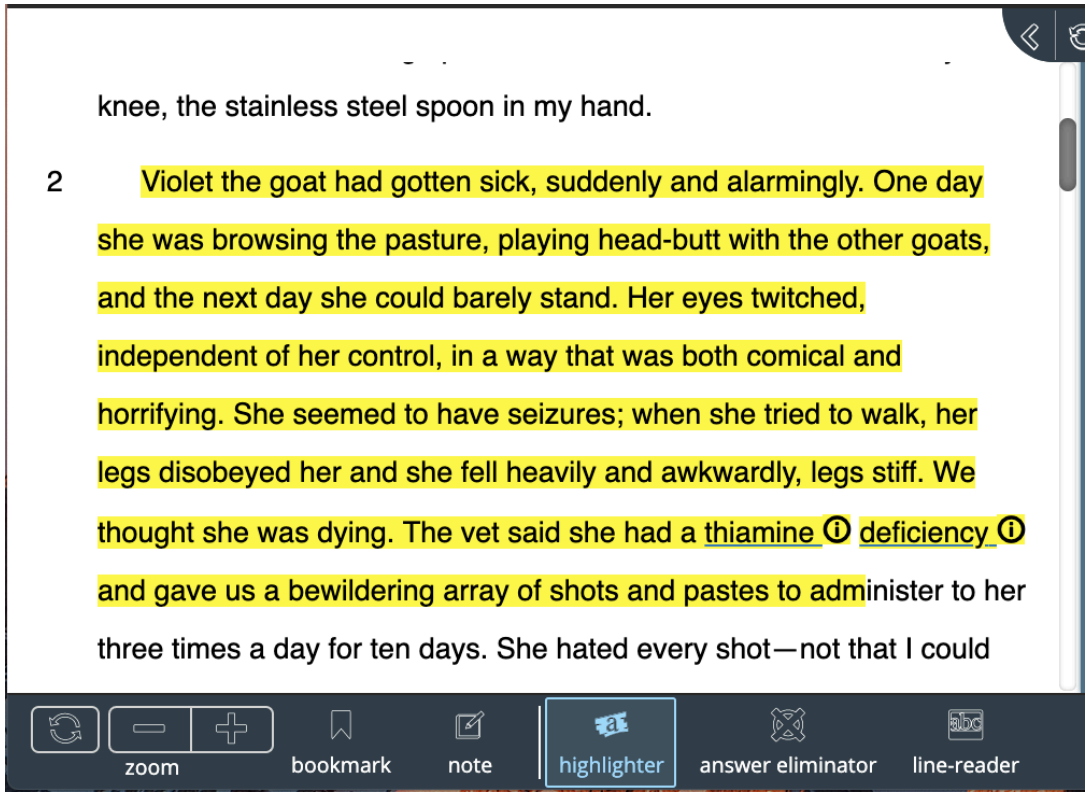
Findings

- Different samples lead to different item parameter estimates
- The difference of item mean P+ is not affected much by the item location in the test
- Easier items are more affected

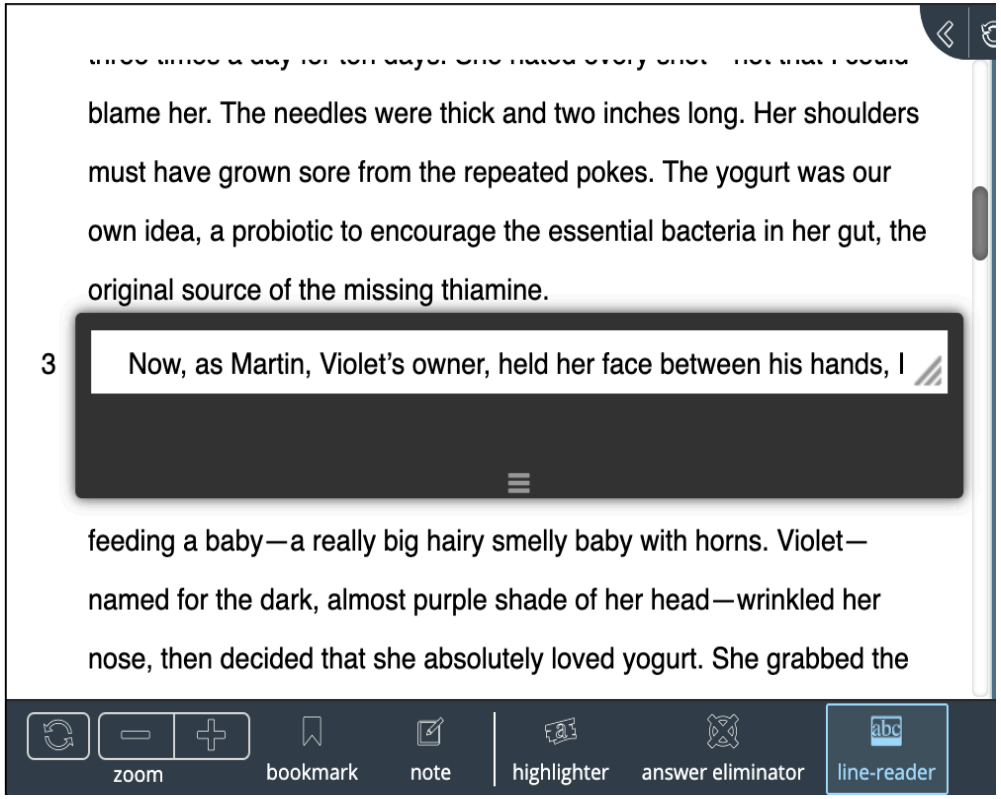
Based on data from ETS Winsight assessment

Example 2: process data for group difference

Highlighter tool

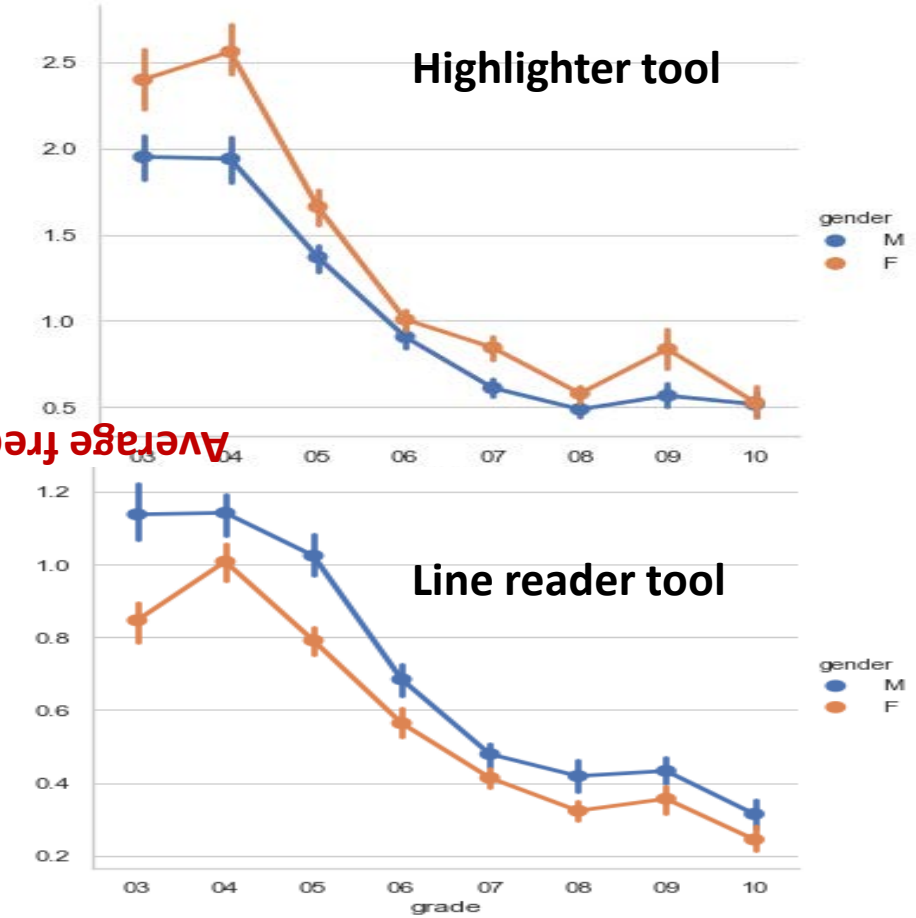


Line reader tool



From ETS Winsight Assessment Sampler

Example 2: process data for group difference



Observations:

- Girls tend to use more highlighter tool
- Boys tend to use more line reader tool
- Decreasing trend of tool usage

How will the tool usage affect the performance? What should be the right tools to be provided?

Summary

- Leveraging the process data can help to improve psychometric quality, discover new patterns and trends, inform better teaching and learning, and form new hypothesis to extend our knowledge horizon.
- Unleashing the full power of process data in large scale assessments requires the proper methods and tools. I hope the glassPy example can help you to think about your ways of dealing with process data.
- Traditional psychometric methods need to be extended to include methods from data science and machine learning to handle the increasingly complex data
- A forthcoming new book, Computational Psychometrics, edited by Alina von Davier, Robert Mislevy and Jiangang Hao will provide psychometrics researchers with a curated list of tools for addressing the challenges from complex data.

Email: jhao@ets.org