Mini-Project #1: Predicting the difficulty of a test item

Overview

This mini-project will give you an opportunity to explore the psychometric properties of several different educational assessments that have been developed to measure "data visualization literacy." You'll work with a dataset containing responses from participants who answered questions about various data visualizations, with the goal of understanding what makes some test items more challenging than others.

These data were collected in a recent experiment conducted in Dr. Fan's lab where U.S. adult participants were recruited online to complete a subset of the items from five assessments, summarized below. Note that these data are not yet publicly available, so it's a special opportunity to dig into real research data in advance of when they would appear in a research publication.

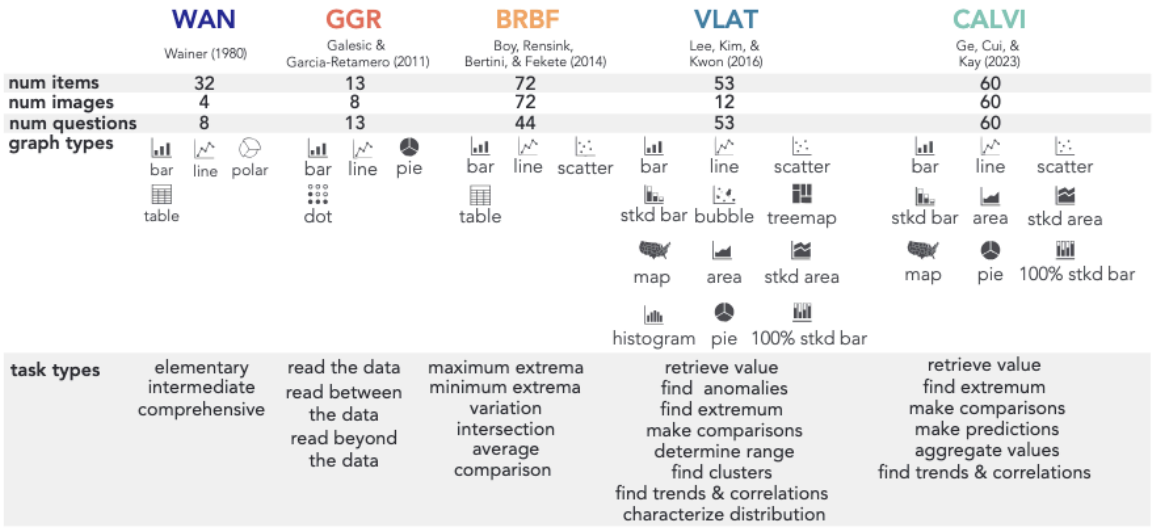
Learning Objectives

- Develop predictive models for item-level difficulty based on various features
- Evaluate model performance using appropriate metrics
- Interpret findings with respect to the assessment of data visualization literacy
- Practice communicating your findings in a 3-minute lightning talk (3-5 slides) and a brief written report (2-page extended abstract following ACM *Learning@Scale* template)

Dataset Description

The dataset consists of responses from a large sample of U.S. adults (N=426) who completed items from five widely-used tests of data visualization literacy:

1. **WAN** (Wainer, 1980): Elementary-level assessment with bar charts, line charts, polar plots, and tables
2. **GGR** (Galesic & Garcia-Retamero, 2011): Assessment with bar plots, line plots, icon arrays, and pie charts
3. **BRBF** (Boy, Rensink, Bertini, Fekete, 2014): Assessment measuring influence of data and visual properties across different graph types
4. **VLAT** (Lee, Kim, & Kwon, 2016): Assessment with 12 different chart types using real-world data
5. **CALVI** (Ge, Cui, & Kay, 2023): Assessment containing standard items and items designed to mislead users



Each row in the dataset represents a single participant's response to a specific test item, with the following columns:

Column Name	Description
`participant_id`	String. Anonymized identifier for each participant
`question_text`	String. The full text of the question asked
`image_url`	String. Link to the data visualization image shown to the participant
`item_id`	String. Unique numeric identifier for each test item
`test_name`	String. Name of the test (WAN, GGR, BRBF, VLAT, or CALVI)
`graph_type`	String. Type of visualization (e.g., bar, line, scatter, pie, etc.)
`task_type_merged`	String. Categorization into value identification, arithmetic computation, or statistical inference
`task_type_original`	String. Original task categorization from the source test
`table_presentation_format`	Boolean. Whether the item used a table or plot to present data
`misleading_item`	Boolean. Whether the item was designed to mislead.
`participant_response`	String. The full text string containing the response provided by the participant
`correct_response`	String. The full text string containing the correct answer to the item

`possible_responses`	Array. For multiple-choice questions, this array contains both the correct response and participant's response.

Goal Specification

You will be given data for 80% of the items across all five of the above assessments. **Your goal is to develop models that can predict the difficulty of the remaining 20% of items.**

The teaching team will not be providing instruction in any particular techniques from the broader data science toolkit. You will generally be expected to have enough background knowledge (and/or willingness to learn in a self-directed manner) about methods for wrangling, exploring, visualizing, and modeling data to be able to apply that knowledge to these datasets. In addition, you are generally expected to have enough experience with scientific writing to be able to prepare the project reports in a mostly independent manner. That said, the teaching team will generally be available during the hands-on portion of each class period to answer questions and provide other forms of individualized support.

Project Timeline

Phase 1: Exploring and visualizing the data

In this phase, you'll aim to:

- Load and examine the data (using any Python/R-based data analysis environment you are comfortable with)
- Perform any necessary data wrangling
- Devise a reasonable metric for item difficulty. By default, we suggest defining the difficulty of a test item as the proportion of participants who respond to that item correctly in our sample.
- Create exploratory data visualizations to explore variation in average item difficulty
- Explore relationships between test name, graph type, task type, and item difficulty

Phase 2: Defining & evaluating statistical models

In this phase, you'll aim to:

- Formulate 2-3 targeted research questions in natural language that can be answered with this dataset
- For each research question, devise various alternative modeling approaches that could be used to predict item difficulty. By default, we expect most students to rely on multiple linear regression, but other methods can be used instead.

- Devise a reasonable measure of model performance and use this metric to compare these models to each other.
- Ensure that any predictive models you devise can be applied to novel items in the test split for this dataset.

Phase 3: Lightning talks & preparation of project report

In this phase, you'll be:

- Preparing a 3-minute lightning talk presentation on your analysis strategy & findings (in class)
- Submit a 2-page extended abstract (due on Friday of Week 4 by 9PM on Canvas). Make sure that you have a link embedded in this extended abstract to a GitHub code repository that contains all of the analysis code needed to reproduce the figures and quantitative results reported in your extended abstract.
    - If your code is formatted appropriately, we will later evaluate your models on the test split from this dataset & share results with the whole class later in the quarter.

Project Deliverables

Oral presentation (3 minutes max)

Your presentation should be accompanied by slides and be no longer than 3 minutes. Successful presentations will:

- Pose the research question in plain, accessible language
- Outline your analytical approach
- Present your key findings with well designed visualizations
- Discuss implications for assessment of data visualization literacy

Extended abstract (2 pages max)

Your extended abstract should follow the ACM *Learning@Scale* LaTeX template provided in the syllabus. It should contain the following sections:

- Abstract (150 words): Provide overview of the project goals, approach, findings in plain and concise language
- Introduction: Describe the problem and its significance in educational assessment
- Methods: Summarize your analytical approach and modeling techniques
- Results: Present your key findings accompanied by well designed visualizations
- Discussion: Interpret your results, discuss limitations, and suggest implications

Analysis code (private GitHub repository)

We expect most analysis code to take the form of either a Jupyter or RmD computational notebook. It should be well organized, with clear Markdown headers that state the research question in natural language and well documented code.

Evaluation Criteria

Your grade for Mini-Project #1 will be based on a combination of the technical soundness of your approach, as well as clarity of communication in both the oral presentation and written report.

Getting Started

Please download the dataset using this URL: https://data-visualization-benchmark.s3.us-west-2.amazonaws.com/vt-fusion/psych_139_mini_project_1_split_80_responses.csv. Enjoy!

Getting Help

The instructors will be available to answer any questions during the hands-on period during each class period. In addition, please feel free to use Slack (e.g., the course channel) to reach out to other members of the class for help (but please refrain from duplicating one another's work). The point of this class is for *you* to learn & develop these skills!