NAEP RESEARCH & DEVELOPMENT (R&D) VIRTUAL SHOWCASE

Highlighting work funded under NAEP Cooperative Agreements

http://naep-research.airprojects.org/

Tuesday, December 8, 2020

Questions and Answers

Each presentation in this NAEP R&D Showcase included a brief moderated Q&A with expert panelists Rochelle Michel (Curriculum Associates), Sarah Theule Lubienski (Indiana University), and David Thissen (University of North Carolina at Chapel Hill). Below are the responses only to questions posed by our external audience during the webinar.

Automated Scoring of NAEP Constructed-Response Items

Presenter(s): Dr. Derek Justice and Dr. Corey Palermo, Measurement Incorporated

Q1. What was the length of the responses to the items chosen by Measurement Incorporated for automated scoring? Did they train their items based on one or two scorers? If two, did they look at the average score? How did they choose a score to train on?

A1. The mean response length was 26 words across items. We were limited to the available data, which included the first score and a second score (where available), but no resolutions or otherwise "true scores." We used matched R1/R2 scores, where available; for most responses, we had a single score.

Q2. I'm curious about whether failed items could be assigned draft scores that a human would check, and whether would that also save costs?

A2. Reducing costs would require reducing the volume of human scoring. If humans had to review all machine scores for the failed items, it would be unlikely to reduce the human scoring time and thus costs. However, one option we touched on in the presentation would be to add routing based on machine "confidence." In such a case, the automated scores could likely be assigned the responses that best matched the training responses (without human review), and the subset of responses that least matched the training responses could be routed to humans for scoring (or review of automated scores).

Q3. What language/platform did you use to generate your machine learning metrics?

A3. We used c#, python, and native code to generate machine learning metrics.

TestBuilder: A Test Design Assisting Tool

Presenter(s): Dr. Igor Griva, IAGriva Consulting

Q1. I'm interested in the size and scale of the application. How many models have you built in one build cycle, and how many items are in the pool?

A1. It is a large-scale application, and it can handle pools of up to several hundred items. One can afford running the code for a few minutes, if needed. The user is anyone who has a pool of items described by their characteristics (i.e., metadata) and who needs to build blocks of items. At the moment, there is only one model, but it can be adjusted to specific needs.

Q2. Are you using a linear programming optimization solver/algorithm on the back end?

A2. Yes, I am currently using CPLEX on the back end.

Q3. Could you describe the user a little more? Test makers at companies, teachers, or others in the education field?

A3. The user could be anyone who has a pool of items described by their characteristics (i.e., metadata) and is interested in building goal-oriented tests to assess examinees' knowledge. That could be teachers or test makers at companies. One of the central characteristics of the potential user is the availability of a well-described pool of items. Whoever has item pools can use the tool.

Bayesian Probabilistic Forecasting with State NAEP Data

Presenter(s): Dr. David Kaplan, Professor, University of Wisconsin-Madison

Q1. Would the revenue differences reflect the economic conditions of the state?

A1. Yes, they would. And, although it would have been nice to have other state-level economic information, nothing else was available to us. However, a serious forecasting model would need such information. In any case, our approach was just an example.

Enhancing the Validity of NAEP Interactive Computer Tasks through Detection of Student (Dis)engagement and Augmentation

Presenter(s): Dr. Yigal Rosen and Ms. Kristin Stoeffler

Q1. How did you decide on the 5th percentile as the criterion for disengagement?

A1. We decided that the top 5% of the fastest responses among all students who responded to an item would be considered disengaged, as they are essentially responding too quickly to reach the end of the task or the assessment session.

Q2. How are reading-level challenges being addressed or considered?

A2. To keep up with the NAEP standards for reading level, we maintained as much of the current text as possible and aligned the new text with the current reading level of the items.