Title: Crowdsourcing the Rating of Open-Response Mathematics Questions

Author(s):

Nathan VanHoudnos  nmv@cmu.edu  Carnegie Mellon University
Lindsey Smith  LSmith@propelschools.org  Propel EAST Charter School
Jamie Callan  callan@cmu.edu  Carnegie Mellon University
Laura Dabbish  dabbish@cmu.edu  Carnegie Mellon University
Brian Junker  brian@stat.cmu.edu  Carnegie Mellon University
Abstract Body

Limit 4 pages single spaced.

Background / Context:
Description of prior research and its intellectual context.

Although open response questions are a more authentic form of assessment than multiple choice or short answer questions, open response questions take more time to score.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

This study presents a technique to reduce a teacher’s grading burden by off-loading the scoring of open response questions. We follow the approach of the computer supported peer review literature, but instead of using non-experts who are students embedded in the context of the classroom, we use non-experts from an online labor market who are independent of the classroom context.

Setting:
Description of the research location.
(May not be applicable for Methods submissions)

The online labor market we have chosen is Amazon Mechanical Turk (AMT). AMT is widely used in the machine learning community to recruit large numbers of anonymous workers to perform simple tasks such as determining if a photo contains a face. This general idea of taking a small task and distributing it to many people is called crowdsourcing.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)

We give the workers on AMT (Turkers) a more complex task. Specifically, we ask them to rate anonymized, hand-written, 3rd-8th grade mathematics open-response questions. These responses were collected from an administration of a 4Sight exam to students who attend a charter school in western Pennsylvania.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
(May not be applicable for Methods submissions)

The Turkers rated the student responses using a web-based rubric designed to offer appropriate scaffolding. Each Turker could rate as many or as few student responses as they wished. Turker ratings were collected until each student response had a set number (from 5 to 30) independent ratings from Turkers. These independent ratings were then combined into a single score for each question and then compared against the score the teacher had already assigned.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.
Previous work in crowdsourcing has not considered an application to education. Previous work in education research has not considered crowdsourcing methods.

**Statistical, Measurement, or Econometric Model:**  
*Description of the proposed new methods or novel applications of existing methods.*

We propose a simple normal model and derive several estimators that account for Turker bias.

**Usefulness / Applicability of Method:**  
*Demonstration of the usefulness of the proposed methods using hypothetical or real data.*

We calculate a reliability (.75 weighted kappa) and validity (0.6 out of 4.0 points mean absolute error) for the aggregated Turker ratings. This performance is consistent across all estimators, and was achieved for approximately the cost of feeding pizza to the volunteer teachers while they made the gold-standard.

**Conclusions:**  
*Description of conclusions, recommendations, and limitations based on findings.*

This study is a proof-of-concept only. Although we have demonstrated reasonable performance for our dataset, it is unclear under what conditions this performance will continue to hold.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.
Appendix B. Tables and Figures

Not included in page count.