Can the Internet grade math? Crowdsourcing a complex scoring task and picking the optimal crowd size

Nathan Van Houdnos
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/statistics

Part of the Statistics and Probability Commons
Can the Internet grade math? Crowdsourcing a complex scoring task and picking the optimal crowd size*

Nathan VanHoudnos†

Heinz First Paper / Advanced Data Analysis Paper / Integrative Interdisciplinary Project

September 20, 2011

Abstract

This paper presents crowdsourcing as a novel approach to reducing the grading burden of constructed response assessments. We find that the average rating of 10 workers from a commercial crowdsourcing service can grade student work cheaply ($0.35 per student response) and reliably (Pearson correlation with the teacher’s gold-standard scores, $\rho = 0.86 \pm .04$ for Grade 3 to $\rho = .72 \pm .07$ for Grade 8). The specific context of our proof-of-concept dataset is 3rd-8th grade constructed response math questions. A secondary contribution of the paper is the development of a novel subsampling procedure, which allows a large data-collection experiment to be split into many smaller pseudo-experiments in such a way as to respect within-worker and between-worker variance. The subsampling procedure plays a key role in our calculation that the average of 10 workers’ scores suffices to produce reliable scores.

1 Introduction

Teachers are busy. Time use studies suggest the average teacher needs over 50 hours per week to complete his or her work (Drago, Caplan, Constanza, & Brubaker, 1999; NEA Research, 2010). Teachers say spending this additional time beyond their contractually obligated 37 hours per week is stressful. When asked

What in your present position as a teacher HINDERS YOU MOST from providing the best service of which you are capable?

and given a blank space with which to respond, teachers consistently answer

Heavy work-load.


Assessing students is part of what keeps teachers busy; they devote between one-third to one-half of their workweek to assessment related activities. Specifically, just scoring student work and recording it consumes 10% to 15% of their workweek (Stiggins & Conklin, 1992). Since that is a significant portion of their week, between 5.7 and 8.7 hours\(^1\), reducing the time it takes to score student work is one way to help teachers be less busy.

Teachers devote such a large share of their workweek to assessment because they believe that it is valuable for “planning instruction, shaping instruction as it unfolds, gauging student achievement, and evaluating

---

*This work was done in collaboration and cooperation with Lindsey Smith, Propel EAST, Pittsburgh, PA to satisfy the requirements of the Heinz First Paper, the Statistics Advanced Data Analysis Project, and the PIER Integrative Interdisciplinary Project. The project advisory committee was Brian Junker, Jamie Callan and Laura Dabbish of Carnegie Mellon University, Pittsburgh, PA.

†Supported by PIER (Program for Interdisciplinary Education Research), IES Grant R305B090023.

\(^1\)Stiggins and Conklin (1992) based their figures on eight case studies of high school teachers over twenty years ago. We report their estimates because other time use literature either (1) combines time spent grading with lesson planning (Drago et al., 1999; Stoddard & Kuhn, 2008; NEA Research, 2010), (2) focuses on the tasks the students perform (Sanford & Evertson, 1983; Roth, Brooks-Gunn, Linver, & Hofferth, 2003), or (3) studies teachers in non-core subjects (Torres, Ulmer, & Aschenbrener, 2007).
curriculum” (Young & Kim, 2010, p. 5). The most valuable kinds of assessment for these purposes are open-ended in nature, but those kinds of assessment take the most time to grade (Sadler & Good, 2006). Multiple-choice and fixed answer questions can be quickly graded, but they hinder the assessment of higher-order thinking processes because their fixed format encourages the assessment of low-level cognitive tasks (Senk, Beckmann, & Thompson, 1997). In contrast, constructed response questions take longer to grade, but they give teachers better insight into student understanding because their open nature allows for the students to express competence in a realistic setting (Jonsson & Svingby, 2007).

If a teacher’s busy schedule effectively forces their students to take only assessments that the teacher can quickly grade, then the student is denied the full benefit of their teacher’s expertise. Low-level assessments (e.g. factual recall, calculation) cannot tell the teacher about the student’s higher-order abilities (e.g. reasoning, problem solving) (Lin, 2006; Young & Kim, 2010). Therefore, if a teacher’s time constraints restrict them to easy-to-grade low-level assessments for the routine measurement of their students, then the teacher may not be able to prepare his or her students for higher-order assessments. These higher-order assessments could be the end of unit exams that the teacher writes or the constructed response items on a high-stakes end of year exam that the state administers. In either case, an over-worked teacher may be surprised to find what their students truly know.

The goal of this paper is to reduce a teacher’s cost for assigning high-level assessments to students. Both high school math teachers (Senk et al., 1997) and elementary school teachers (Jackson, 2009) have been observed to rely primarily on low-level multiple choice questions for assessing students. If we can lower the cost of assigning high-level assessments, then we would expect teachers to assign more of them. Since the primary cost difference between assigning a low-level assessment and a high-level assessment is that high-level assessments require more time to score (Sadler & Good, 2006), our approach focuses on scoring high-level assessments for the teacher. We specifically present a proof-of-concept that constructed response math questions can be scored by anonymous people on the Internet cheaply ($0.35 per student response) and reliably (Pearson correlation with the teacher’s gold-standard scores, \( \rho = 0.86 \pm 0.04 \) for Grade 3 to \( \rho = 0.72 \pm 0.07 \) for Grade 8).

We develop the proof-of-concept for this novel approach as follows. We begin by reviewing other approaches for reducing the grading burden and introduce crowdsourcing, the general method underlying our novel approach. We then give the motivating example of a charter school that asks its teachers to volunteer their time after school to hand-score a school-wide exam, and then we use that very specific example throughout the paper to demonstrate the general ideas of our novel approach. We do this in two parts, where each part answers a related research question:

- Can anonymous people on the Internet score student work? (Yes)
- How many of these anonymous people must score a student response before the average of their scores agrees with a teacher? (About 10)

We then conclude with a brief discussion.

## 2 Related work on reducing the grading burden

The existing solutions to reducing the grading burden fall into two broad categories. The first, which we do not discuss, is increasing the efficiency of how the teacher scores. These are useful methods, but even efficient teachers have limited time. The second type of solutions remove the scoring task from the teacher; they ask someone or something else to score the student work. We categorize these offloading approaches into four types depending on who or what is scoring the student work. The categories are computer automated scoring, peer-scoring, self-scoring, and hiring teaching assistants.

Computer automated scoring comes in two kinds. The first kind are systems that score either multiple choice questions or fixed answer questions. The system knows which response is correct because the student made a particular kind of mark with a pencil, or clicked a particular radio button in a web browser, or entered a specific number (15.3 m/s) in a text box. As discussed, this kind of system cannot score high-level open format assessments, so we do not consider it further. The second kind of computer automated scoring builds models to predict the score that an expert human rater would assign. For example, a computer can “read” an essay, extract the low-level syntax features, fit a statistical model, and output a predicition of
the score an expert human rater would have assigned. The computer does not understand what it is doing; it is simply returning the most likely result from the model (Bennett, 2011). In practice, this works fairly well for structured essays such as GMAT scoring by e-Rater (Burstein, Leacock, & Swartz, 2001; Attali & Burstein, 2005), but it works less well for creative essays or poems (Bennett, 2011). Beyond expensive computer-based systems that guide student responses into pre-set structures, computer scoring does not work for mathematics (Bennett, 2011). Since computer automated scoring cannot currently help a teacher with the kind of open pencil-and-paper assessment shown in Figure 1, it is of limited use for reducing the grading burden.

30. Nadia is learning about functions in her math class. She was given the assignment below for homework. Complete the assignment.

A. The function table below shows numbers that are related by a rule.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>48</td>
</tr>
</tbody>
</table>

Write the missing number in the empty space in the table.

B. Describe the rule for the function table in part A.

The rule for part A’s function table is to multiply x by 12 to get y.

C. Make a new rule with a different operation. Use the rule to create a new function table with three pairs of numbers.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

D. Describe the rule for the function table you created in part C.

The rule for my function table is to get y by getting the next multiple of x.

Figure 1: An example of a student response to a constructed response question taken from the 4Sight Grade 6 mathematics exam.

Peer-scoring uses the other students in the classroom to score a given student’s work. This can be as simple as passing papers to the student on the left, or as complicated as computer systems designed to scaffold a peer review process. In either case, since the students score in parallel, the students receive feedback quickly.
This helps the students on writing assignments, since the fast-feedback allows the students more time to write more drafts, instead of waiting for the teacher to give them feedback. In addition, at the post-secondary level, there is some evidence that critiquing another student’s work as part of a computer-aided peer review process improves the writing of the reviewing student (Goldin & Ashley, 2010; Cho, Schunn, & Kwon, 2010). However, peer-scoring suffers from two major flaws. First, students perceive peer-scoring negatively. They do not believe that other students should hold the power to determine grades (Cho, Schunn, & Wilson, 2006; Sadler & Good, 2006; Kaufman & Schunn, 2010). Second, for elementary and secondary education, peer-scoring requires class time that may be better spent on other higher-value instructional activities. The efficacy of peer-scoring as a learning tool for this age group is unproven. To our knowledge, no published studies examine whether peer-scoring improves writing for elementary or secondary students. The single study to examine peer-scoring in this age group concluded that it does not help students learn middle school science (Sadler & Good, 2006). Therefore, since peer-scoring is both unpopular with students and unproven as a method for helping them learn, it is of limited use for reducing the grading burden.

Self-scoring asks the student who completed the assignment to score his or her own work. To our knowledge, only Sadler and Good (2006) have studied self-scoring in K-12 education. They report that self-scoring improved the performance of middle school science students, but that the self-grades were biased. Low performing students consistently rated themselves higher than the teacher would. However, Sadler and Good argue that even biased scores still will save teachers time because the teacher must only focus effort on the outliers of the reported scores. We agree that a biased score is better than requiring the teacher to generate all of the scores, but we argue that self-scoring takes too much classroom time. Sadler and Good report that self-scoring required extensively training the students to apply scoring rules correctly, and it required building trust so that students did not attempt to falsify their scores. In addition, students felt uncomfortable scoring their own work. Therefore, since the weight of a single study reporting learning gains does not outweigh the large amount of classroom time self-scoring consumes or student’s negative attitudes towards it, self-scoring is of limited use for reducing the grading burden.

Hiring teaching assistants would be ideal for reducing the grading burden if they were not so expensive. Teaching assistants have enough expertise in the area they are teaching to both guide students and score students’ work with minimal direction. This expertise allows them to give helpful feedback to the students as well as summarizing student performance to the main instructor. This is common at the post-secondary level, but at the secondary and elementary levels, budgets are too tight to allow for this expense.

In summary, elementary and secondary teachers must pay too much in either time or money to offload the grading burden of high-level assessments. Computer automated scoring systems are expensive to purchase and too restrictive in the assignments they can score. Peer-scoring and self-scoring can score arbitrary high-level assessments, but they are biased, require a large investment of classroom time, and are viewed negatively by students. Teaching assistants are accurate and able to offer feedback, but K-12 schools cannot afford them. In the next section we introduce a new approach that seeks to emulate the accuracy of teaching assistants at a fraction of their cost.

3 A new approach to reducing the grading burden

Teaching assistants are expensive partly because they are experts. In subjects like math and science, they know the course material well enough to guide students through solving problems, and they can write their own rubrics to assess student performance. This is more expertise than is needed to simply grade student work with a pre-defined rubric. For example, Sadler and Good (2006) found that training seventh grade students on a teacher-developed rubric and scaffolding them throughout the rating process yielded a high correlation ($\rho = 0.94$) between the student generated scores and teacher’s scores. The seventh grade students in the study were likely not able to serve as teaching assistants to each other, but they were able to score each other’s work. Although they were not experts in scoring, students performed at the level of experts because they were specifically guided in that particular task.

This general idea of replacing experts with non-experts is studied extensively in the crowdsourcing literature. Crowdsourcing takes a complex task commonly performed by an expert, breaks it up into smaller pieces, and then distributes those pieces to a group (crowd) of workers. The crowdsourcing platform guides workers through their portion of the task, and when they finish, the platform combines the results, delivers
those results to the user who posted the work, and then pays the workers after the user has verified the quality of the results (Law, 2011). This approach lowers costs because it takes a few expensive experts completing a single complex task and replaces them with many inexpensive non-experts carefully guided through many smaller tasks.

In crowdsourcing, the choice of crowd must be balanced with the guidance offered to the crowd. For example, a professor handing a set of exams to a group of teaching assistants does not need to give them step-by-step guidance on grading, but a teacher asking students to pass their papers to the left must give the students explicit, clear guidance on how to grade their peer’s work. If the teacher chooses to hand the student’s exams to non-experts outside the context of the classroom, the teacher must give careful thought to who those non-experts are and how to guide them.

The novel idea this paper explores is replacing the crowd of peers within the classroom with a crowd of anonymous people from the Internet so that the teacher receives the benefit of an external teaching assistant’s score with no cost in classroom time and a fraction of the cost of a teaching assistant’s time. We explore this general substitution by using a commercial crowdsourcing service to solve a problem that a local charter school faces with the scoring of a constructed response exam. Specifically, we use Amazon Mechanical Turk\(^2\) (AMT) to score Grade 3, Grade 6, and Grade 8 constructed response math questions from the Success for All Foundation’s 4Sight exams\(^3\) that our partner charter school currently grades by asking teachers to volunteer after school. The next section of the paper addresses if this approach is possible and the following section determines how much it will cost to achieve acceptable results consistently.

4 Can anonymous people on the Internet score student work?

This section explains the careful thought required to scaffold a crowd drawn from Amazon Mechanical Turk in performing a complex constructed response scoring task. We begin with a description of the general context of Amazon Mechanical Turk and then explain the context of our partner charter school. We then move to the key part of crowdsourcing, which is designing the scaffolding the non-experts use. We then move to details of the rating study the teachers participated in and contrast them with the details of the rating study conducted on AMT so that the teachers and crowd can be directly compared. We then close the section by comparing the teacher scores, a crowdsourced rating study where five workers rated each piece of student work, and a second crowdsourced study where thirty workers did so.

4.1 Amazon Mechanical Turk: The source of our non-experts

Amazon Mechanical Turk is the largest and oldest general purpose crowdsourcing service. The workers on AMT are an international group, approximately 45% are from the United States, 34% are from India, and the remainder are from elsewhere in the world. Since approximately 90% of them have at least completed some college, we can assume that the majority are qualified to grade math questions from grades 3-8 (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010; Paolacci, Chandler, & Ipeirotis, 2010). AMT has its own vernacular. Four terms are particularly important:

- **Turkers** are the people who work on AMT;
- **Requesters** are the people or organizations who ask the Turkers to perform work;
- A HIT, or Human Intelligence Task, is the basic unit of work that a Requester offers to the pool of Turkers; and
- an Assignment is a collection of HITs that the Requester assembles together to accomplish to his or her specific task.

To use these terms in a simple example, if Amazon.com wants to use AMT to verify that all of the books in their current-event category are, in fact, books about current events, then Amazon.com would be the Requester. An individual HIT might be defined as asking a Turker to check if a single book describes a

\(^2\)http://www.mturk.com

\(^3\)http://www.successforall.net/
current event, and then the whole Assignment would be a list of all the books currently classified as current event books.

To explain the common use of AMT, we give a more involved example. Consider the case of a Requester who has a large number of photographs (10,000) taken by tourists that may or may not contain a corporate logo. The goal is two-part. First, to determine the subset of photos that contain logos and second, to assign short descriptive labels to all of the images. Since the two goals do not depend on each other, the Requester chooses to split the work into two separate Assignments, each with their own type of HIT. The first type of HIT asks the Turkers to answer a Yes/No question (“Does this image contain a corporate logo?”) for ten photographs. The second HIT would be very similar to the first, but instead of asking a Yes/No question, a blank text box would be displayed under each image so that the Turker could type in an appropriate label. Given the more open nature of the second HIT, the Requester must pay close attention to the instructions given to the Turkers. Otherwise subsets of Turkers may interpret the directions in one way while a different subset may interpret the directions in different way (Welinder, Branson, Belongie, & Perona, 2010). It is the responsibility of the Requester to ask specifically what is desired from the crowd so that the Turkers are not misled and reliable results are obtained (Paolacci et al., 2010; Hullman, 2011).

The Requester should give careful thought to the grain-size of the HIT, the amount of work a given HIT contains. Depending on the dimensions of the photos and the typical resolution of a Turker’s computer display, the Requester could choose between a single photograph or tens of photographs to comprise a single HIT. Too many photographs, and the Turkers may see the task as overly burdensome, but too few, and the Turkers may find the overhead delay to get the next set of photographs negatively impact their efficiency (Chilton, Horton, & Miller, 2010). Once the Requester chooses the grain-size, the Requester then determines the expected amount of time to complete a single HIT. This allows for a calculation of a fair payment to offer the Turkers in a given Assignment. If the pay is too low, not enough Turkers will select the HITs to complete the Assignment of all 10,000 photographs in the dataset (Mason & Watts, 2009), but if the pay is too high, the Assignment becomes more attractive to Turkers who wish to game the system (Downs, Holbrook, Sheng, & Cranor, 2010).

Since the Requester does not know the quality of the Turkers beforehand, it is common to have each HIT completed by multiple Turkers. If, for example, within the same HIT, three Turkers classify each photo as either including a logo or not including a logo, then at least two of the three Turkers will agree. This majority agreement is a common way to mitigate the effects of Turkers who attempt to game the system. The “good” Turkers are expected to outnumber the “bad” Turkers, so as the number of independent Turkers who classify each photograph increases, the simple majority vote will tend to the correct answer. The number of Turkers needed before the majority vote is of acceptable quality is dependent on the complexity of the specific HITs. There is much active research in methods that replace the majority vote with a different statistic that achieves an acceptable quality with fewer Turkers required and therefore lowers the cost to accomplish the task (for further details, see the review in Raykar et al., 2010).

Ensuring that the Turkers are working honestly is relatively easy for the simple HITs in this example. One approach is to directly compare the Turkers performance on a small gold-standard subset of photographs that are known to either contain, or not contain, a logo. If a Turker mis-classifies too many of the known photographs, that Turker would simply not be paid (Shaw, Horton, & Chen, 2011). In a similar way, if the Turker did not specify enough acceptable labels for a gold-standard image, then the Turker would not be paid. When using this approach, it is important that every Turker perform enough tasks so that the Requester can accurately judge the quality of a given Turker. If, for example, every Turker merely classified one image, then worker quality could only be determined for Turkers who were assigned to the gold-standard data.

Now consider a slight modification to the example. If the Requester is now interested in the interaction between a Turker first looking for a corporate logo in an image and then giving a short descriptive label for the image, the task would need to be redesigned. A single HIT must now accommodate both tasks, as that is the only way to ensure that a Turker would perform both the classification and labeling tasks for the same image. Combining the two tasks into a single HIT may require a reduction in the grain-size of the HIT to accommodate the additional cognitive load of switching between classification and labeling tasks. It may also require a change in payment offered, as the task may now take more time.

In summary, AMT can solve complex tasks if a Requester can break the task into small, easily digestible HITs that a single Turker can solve individually. If so, then collecting the HITs into an Assignment will allow
the Turkers to solve the task in aggregate. In designing these HITs, Requesters must give careful thought to the ability of a single Turker to complete the HIT, but also to the HIT’s curb-appeal so that it can successfully attract the attention of the Turkers when compared against other HITs in the marketplace. Otherwise too few Turkers will choose the HITs in the Assignment, and the Assignment will remain unfinished. Requesters must also think about how to validate the responses the Turkers give, since some Turkers attempt to game the system.

4.2 Propel EAST: The source of our experts

We collaborated with Propel EAST, a charter school in Turtle Creek, Pennsylvania, to give their teachers more time by offloading the scoring of student work. Propel currently asks teachers to volunteer their time after school to score a school-wide exam. The goal is to apply crowdsourcing so it is unnecessary to ask the teachers to volunteer. Instead the teachers could use that extra time to analyze the results of the school-wide exam at a deeper level.

Propel EAST is not representative of Pennsylvania schools. It differs in several respects. Propel EAST is one of several Propel Schools\(^4\), a collection of schools centrally managed by a not-for-profit organization. Since the state of Pennsylvania does not manage Propel’s day to day affairs, Propel has greater control over the hiring and dismissal of teachers. It is a relatively new school; it has only been in operation since 2005 and the majority of the teachers are in the beginning of their careers. Although it primarily serves students from low-income families, the school has made Adequate Yearly Progress each year of its operation. This distinction is not shared by many other Pennsylvania schools that serve this demographic.

Propel EAST is convenient for researchers. The school collects a lot of data on its students. A large part of this data arises from third-party exams that predict the school’s performance on the PSSA\(^5\), Pennsylvania’s high-stakes end of year assessment. Propel EAST uses both computer based adaptive assessments like the Measures of Academic Progress (MAP) exam from Northwest Evaluation Association\(^6\) as well as more-traditional pencil-and-paper assessments that mimic the format and structure of the PSSA like the 4Sight exams from the Success for All Foundation\(^7\).

Although machines score the entirety of the MAP exam, only the multiple choice section of the 4Sight exams can be scored by machine. Each 4Sight exam contains two constructed response questions that teachers must score by hand. This is an inconvenience both to the teachers who volunteer to score (the teachers could be spending this time on other tasks) and the teacher whose students are being tested (unlike the MAP exam, the results are not available until the volunteers can gather to process the scores). Propel accepts these inconveniences because the open-ended nature of the constructed response questions is a valuable component of predicting how the students will perform on the PSSA.

Our collaboration with Propel EAST revolved around designing and pilot-testing a drop-in replacement to their current inconvenient practice. To motivate the issues involved, we give a short overview of the grading task the expert teachers perform.

4.2.1 Rubrics: The tool the experts use

The Success for All Foundation provided a rubric specific to each question. Figure 1 on page 3, gave an example of student work for Grade 6 Question 30, Figure 2 displays the rubric to grade that student work. It is the most complex rubric in the dataset provided by Propel EAST. Its proper use requires careful training and attention to detail.

The rubric in Figure 2 is designed for experts to use. First, it contains embedded exceptions that the volunteer teachers must catch. The scoring of parts A, B and D is very simple. The volunteer teacher must only answer a simple Yes/No question: “Was the answer provided by the student correct?” For example, the student work given in Figure 1 received 1 point for part A, 1 point for part B, and 0.5 points for part D. The rubric becomes more complicated for part C. The volunteer teacher must first determine whether the student followed the direction to not use multiplication. Only if the student used a different operation (i.e. addition, subtraction, or division) is the volunteer teacher to determine the quality of the student response.

---

\(^4\)http://www.propelschools.org

\(^5\)http://www.portal.state.pa.us/portal/server.pt/community/pennsylvania_system_of_school_assessment%28pssa%29/8757

\(^6\)http://www.nwea.org

\(^7\)http://www.successforall.net/
on a scale with three levels (i.e. 1.5 points, 1.0 point, or 0.5 points). For the example response in Figure 1, note that although the student constructed a valid function table, the student ignored the specific instruction not to use multiplication. Consequently, the student received zero points for part C. The raw score of this example student was then 2.5 out of 4 points. The last step the volunteer teacher took, which is not shown in the example rubric, was to convert the raw score to a Likert-type scale using Table 1. This final step also required attention to detail because the table uses a non-standard rounding scheme.

The rubric required the volunteer teacher to pay careful attention to its instructions in order to use it reliably. First, the Yes/No questions in parts A and B were worth 1 point while the Yes/No question in part D was worth 0.5 points. Even though the student from Figure 1 did not follow the directions for part C, the student still earned full credit for the explanation of that incorrect answer in part D. Given that the volunteer teacher was invested in the activity of scoring student work, this is not a source of concern for this traditional method of scoring. It is, however, something that must be addressed if crowdsourcing the scoring.

4.3 Interface Design: the key to crowdsourcing

Crowdsourcing takes a complex task commonly performed by an expert, breaks it up into smaller pieces, and then distributes those pieces to a crowd of workers. How to split the Assignment into HITs and display those HITs is key to the success of crowdsourcing. It is the focus of this section.

We define an interface as the implementation of the design decisions of splitting the Assignment into HITs and building a way to display them to the Turkers. We begin with a description of a bad interface that motivates two general lessons. We then describe the lessons so that the resulting interface successfully splits an Assignment into individual HITs, and displays the HITs intuitively so Turkers will successfully complete the assignment.

4.3.1 An example of a bad interface to learn from

We present the first interface we offered to the Turkers as an example of a bad design to learn from. It failed. Not enough Turkers were willing to complete the HITs, even when offered the comparatively high wage of $0.15 per HIT, so the Assignment never completed.

This first interface closely mimicked what the volunteer teachers did. First, the Turkers read the same rubric the volunteer teachers used. After reading the rubric, the Turkers saw an example of student work for the first part of the first question. Since the Turkers had computer monitors instead of physical paper rubrics and physical paper copies of student work, the interface design kept the relevant pieces of information on the screen at the same time. This display pattern gave a piece of student work, then a relevant piece of the paper rubric, and then finally a multiple choice question that implemented the question asked by the rubric. Three student responses were shown in a single HIT so that the Turkers who invested the time to understand the rubric could quickly rate the work of several students at once.

Figure 3 displays a zoomed out screen-shot to give a sense of the interface’s sheer size. The text is much too small to read because each black arrow represents the height of a standard computer display, approximately 750 pixels. The first three arrows span a complete copy of the rubric and the informed consent. The next four and a half arrows span the response from the first student that is interleaved with the rubric. The last half of the last arrow spans the beginning of the second student’s response. Although not shown in the figure, the interface also displayed a third student response. The Turker had to press the
4Sight rubric used to score Figure 1

30. (Test 2) Top Scoring Response

<table>
<thead>
<tr>
<th>Part A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer:</td>
</tr>
<tr>
<td>$x$</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

(1 score point)
1.0 point for correct answer

<table>
<thead>
<tr>
<th>Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation:</td>
</tr>
</tbody>
</table>

(1 score point)
1.0 point for correct explanation of the rule

<table>
<thead>
<tr>
<th>Part C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer:</td>
</tr>
<tr>
<td>$x$</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>90</td>
</tr>
</tbody>
</table>

OR
Other correct function table that is not a multiplication function

(1.5 score points)
1.5 points for all correct values in the function table (that describes an addition, division or subtraction function)

or
1.0 point for enough correct work to follow what is being done (only minor omissions or errors)
0.5 point for not enough correct work to follow what is being done or some correct work only

Note: if the student creates a multiplication function table they cannot receive any points for part C

<table>
<thead>
<tr>
<th>Part D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation:</td>
</tr>
</tbody>
</table>

OR
Other correct explanation

(0.5 score point)
0.5 point for correct explanation of the rule in part C, even if student created a multiplication function

Figure 2: The 4Sight rubric for Grade 6 Question 30. An example of student work for this question was given in Figure 1
This interface failed in at least three ways. First, it gave the Turkers the expert-level task without breaking it down for them or offering scaffolding. This was an error because the crowd from Amazon Mechanical Turk is not a crowd of teaching assistants. They do not have the expertise to grade with the rubrics the teachers used. Second, this interface was ugly and hard to navigate. This was an error because Turkers choose HITs from a large pool. If a HIT is not attractive, Turkers are less likely to choose it and the Assignment may not complete.

4.3.2 Lesson 1: Split the expert-task and scaffold non-experts

The rubrics the volunteer teachers used give minimal direction because they assume that the people who will use them are experts with long attention spans who will receive training in their use. For example, in Figure 2 the part C portion of the rubric merely reminds the teacher with a note at the bottom of the instructions that no points are to be awarded if the student used multiplication. That exception could be missed by a non-expert if the non-expert is overwhelmed with other information.

The easiest way to scaffold the non-expert Turkers is to reduce the grain-size, the amount of work they perform in a given HIT. If, for example, Turkers were only asked to grade part C in Figure 2, then the exception of using multiplication is unlikely to be missed. It is not, in principle, necessary to keep part A, part B, and part C from the same student together in the same HIT. A Requester can construct HITs of arbitrarily small grain-size as long as the specification of the Assignment still covers all of the parts of all of the questions the students answered.

However, HITs that are too small cannot be easily compared with the teachers ratings. The volunteer teachers rated the student work at the question level, so we can only directly compare individual Turkers
with the teachers when the same Turker rates the whole question. Since this proof-of-concept study requires this direct comparison to identify Turker quality, it sets a limit on the allowable grain-size of the HIT. For our case, a whole question must be a single HIT. Generally, the Requester must choose the grain-size to balance the needs of the Turkers with the needs of the problem the Assignment solves.

Since our grain-size for the Turkers is the same as the grain-size for the expert teachers, we must scaffold the Turkers throughout the task. We chose to do this by collapsing the rubric the teachers used into a series of *judgments*, which are the smallest answerable pieces of a rubric, and then combining the judgments within a HIT to calculate the score the Turker indirectly assigned. For example,

In Part A, did the student give the correct answer (144)?

- [ ] Yes
- [ ] No

is a simple judgment. We use multiple judgments to scaffold Turkers through rubric exceptions by hiding the scoring consequences of the exceptions and instead only asking the Turker to find the exceptions. For example, the multiplication exception in part C of Figure 2, can be implemented by asking:

In part C, did the student use an operation other than multiplication to construct their table?

- [ ] Yes
- [ ] No

For part C, rate the quality of the student’s work.

- [ ] Everything is correct.
- [ ] Minor errors are present.
- [ ] There is only some correct work.
- [ ] The answer is wrong.

and then using the answers to these two judgments to calculate the appropriate score for the HIT.

In summary, the first design lesson is to break the Assignment into the smallest possible HITs and scaffold the Turkers throughout the remaining within-HIT judgments. This will give a task-level design that Turkers will be able to complete. The next lesson addresses attracting the Turkers to these split and scaffolded Assignments.

### 4.3.3 Lesson 2: Make the HIT attractive and easy for the Turkers to use

Turkers find Assignments through text-only advertisements on the AMT Marketplace. If a Turker is interested in an Assignment, AMT gives the Turker a preview of a randomly selected HIT from the Assignment. The Turker may then choose to accept the HIT or return to browsing for other Assignments. Consequently, the preview of a HIT is part of the advertisement that draws Turkers to the Assignment.

Therefore, if the HIT does not immediately engage the Turker with clear and interesting work, the Turker is likely to move on to other Assignments. Careful choice of HIT grain-size and scaffolding for within-HIT judgments will bring the HIT close to the expectations of the Turkers, but the HIT must also be easy to use.

Figure 4 displays the interface we designed for the Turkers. We call it the Turker-centric interface and we demonstrate three reasons why it is attractive to the Turkers by contrasting it with the mistakes of the bad-interface previously described from Figure 3.

First, the instructions of the Turker-centric interface are short, a single sentence: “Please grade the following math question.” This brevity communicates the simplicity of the task to a Turker who is previewing the work. In contrast, the bad-interface required three screens of text to explain the task. This was too long; Turkers did not stay to read it.

Second, the Turker-centric interface is minimalist with a plain-background and a simple border to contain the example of student work. This reduces the cognitive load on the Turker by separating the image of the student work from the judgments the Turker must make about that image. This allows the Turker to focus on the task without visual distractions. In contrast, several competing visual elements cluttered the bad-interface. It displayed an image of student work alongside an image of the rubric to grade the student work. The choice to include the information from the rubric as an image diluted the visual impact of the required image of student work. This was visually distracting. In addition, the unnecessary background colors reduced contrast and distracted the eye.
Figure 4: A Turker-centric interface for the 4Sight rubric shown in Figure 2. The omitted student image can be found in Figure 1
Third, the Turker-centric interface subtly informs the Turkers that the Requester will not penalize them for technical errors beyond their control. Each of the judgments in Figure 4 contains a special non-answer option for the Turkers to use if the image of student work does not load correctly, or if a portion of the image is unreadable. At the end of the HIT, there is an open-response comment box for the Turkers to report any problems encountered. These simple measures reassure the Turkers that the Requester has designed the HIT with their interests in mind. In contrast, the bad-interface could not communicate this because its design was too cluttered to emphasize that it contained non-answer options and a comment box at the end of the HIT.

In summary, the second design lesson is that an individual HIT must capture the attention of a previewing Turker with clear and interesting work. Turkers have their choice of HITs from many Requesters, so Requesters must advertise their Assignments well.

### 4.3.4 Summary

Applying the two lessons described above to the expert-level rubric from Figure 2 led to the construction of a Turker-centric interface that succeeded in attracting enough Turkers to complete the entire Assignment. In addition this cleaner design drove down the cost of rating the Assignment. Turkers were willing to score student questions at $0.03 per question with the Turker-centric interface when they refused to score them at $0.05 per question with the bad-interface.

We built a Turker-centric interface for all of the rubrics in the proof-of-concept dataset. We present detailed results of the performance of those Turker-centric interfaces at this end of this section. In the next subsection we explain the Assignment-level choices that Requesters must make after splitting a task into many smaller HITs and the structure of the resulting data.

### 4.4 Organizing HITs into Assignments

We continue the parallel comparison of experts with non-experts as we build on the HIT-level descriptions presented above. Therefore we step back from crowdsourcing to describe how Propel EAST currently scores student work and how much it costs them to do so. We then discuss the rating studies we conduct on AMT.

#### 4.4.1 Pizza parties: How experts complete their Assignment.

Currently Propel East grades the 4Sight exams with their local population of experts, the teachers. Propel distributes the scoring burden across all of its teachers with an open-call to volunteer after school. These after-school scoring sessions are referred to as “pizza parties” because pizza is served as a gesture of thanks for volunteering. The open call usually results in about 7 to 8 teachers for each pizza party. Typically the teachers who attend the pizza party have a strong background in mathematics, but not all of them teach math. Like many of the teachers at the school, most of the teachers at the pizza party are in the first few years of their careers.

The pizza party is lead by a teacher who received extensive training on the 4Sight rubrics. This lead teacher, in turn, trains the volunteer teachers during the pizza party. The training begins with a general overview of how to score the exams (e.g. don’t mark on the exams, feel free to ask questions) and the features common to all the exams (e.g. how to use the specific 4Sight rounding scheme, where to record the students’ scores). After this general orientation, the volunteer teachers are split into grade-levels and they receive a blank copy of the two kinds of questions they will score. The lead teacher then asks the volunteer teachers to independently answer the questions as though taking the test. This gets the volunteer teachers to think about the several ways that such a question could be answered. After completing the questions, each teacher uses the 4Sight rubric to grade his or her own work. At this point, the volunteer teachers ask questions about the rubric among themselves and of the lead teacher.

The scoring begins after all of the teachers understand the rubric and its use to score the exams. The teachers sit together at their grade specific tables and each table splits the student responses amongst its teachers. Each teacher scores their own pile of student responses. No student response is graded by more

---

8Later work, completed after we had designed this interface, demonstrated generally that Turkers demand higher wages for bad interfaces (Toomim, Kriplean, Pörnter, & Landay, 2011).
than one teacher. The teachers can ask questions of their table-mates when it is unclear how to grade a particular student’s response. If the teachers at the table cannot come to a consensus or are unsure about how to proceed, they call over the lead teacher to resolve the issue.

The pizza party outputs a single score for each student. These scores are expensive. A conservative estimate of a Propel teacher’s hourly wage is $13.64. If teachers were compensated for their time, each score would cost at least $0.51 to produce. Since the expert labor is donated, the current expenditure of Propel EAST is only for the pizza that is served as a gesture of thanks. Therefore Propel’s cost is approximately $0.11 per student response.

In summary, expert teachers donate their time to grade the 4Sight exams after school hours. They use an informal, social process where they both score student responses and interact with other teachers. This produces a single expensive score for each student response.

4.4.2 T-Studies: How Turkers complete their Assignment

The organization of the volunteer teachers into a functioning system that took in exams and output a single, high-quality but expensive score for each student response was straightforward. Here our goal is to organize the Turkers into a parallel system that outputs many cheap scores and averages them to obtain a single score that is of comparable quality to the teacher assigned score. Section 4.3, already described much of this organizational work by giving careful attention to the design of the interface the Turkers use. What remains is simply to post the HITs on Amazon Mechanical Turk and collect the Turker responses.

We developed CrowdGrader, a custom web-application to simplify this process of scoring student work on Amazon Mechanical Turk. It does this by providing the appropriate boiler-plate code to interface with AMT and the Turkers so that the Requester can focus on interface design and the output of the rating process.

CrowdGrader scores student work by organizing HITs into an on-line rating study. We define a T-Study as an Assignment that scores all of the student work for a given exam. CrowdGrader requires three pieces of information to define a T-Study on AMT. First, it requires the definition of the rows, which are the set of HITs that the T-Study contains. For example, if 38 students answered 2 questions in an exam and the HITs were defined so that 1 HIT covered 1 question, then the T-Study that rates the exam would have $38 \times 2 \times 1 = 76$ rows.

The second piece of information a T-Study requires is the size of the T-Study, which is the number of independent Turkers who must rate each HIT. For example, if the Requester desired to average 3 independent Turkers into an aggregate score, then the Requester would set the size of the T-Study to 3. CrowdGrader would then collect 3 scores for every row, and the Requester could average across the rows for the desired score.

The final piece of information a T-Study requires is the reward of the T-Study, which is the amount of money offered to a Turker for the successful completion of a HIT. This allows CrowdGrader to determine the cost to rate a T-Study. For example, if a T-Study has 72 rows, a size of 3, and a reward of $0.03, then the T-Study will cost $76 \times 3 \times 0.03 = $6.84 plus fees from AMT.

The output of a T-Study can be thought of as a table. Figure 5 visualizes this table metaphor for a fictitious T-Study with 4 rows and a size of 3. Although the number of rows of the T-Study is the same as the number of rows in the table, the size of the T-Study is not the number of columns in the table. The size of the T-Study is the number of scores each row receives. In Figure 5, it is clear that this table is from a size 3 T-Study because each row has ratings from 3 different Turkers.

We define the persistence of a Turker as the analog to the size of a row, it is the number of ratings that the Turker provided to the T-Study. This is an important quantity, because it is the within-Turker sample size. As the persistence of the Turker increases, more information is gained about that Turker, so the precision of estimates about that Turker’s properties will improve. For example, to quantitatively determine which Turkers should be paid, the Requester must choose a comparison statistic and a cut-off. Turkers with low

---

9Since the length of the Propel school year is 220 days, if we assume that all of the teachers at the pizza party are new teachers who only make $30k per year, and further assume that each of the teachers works a 10 hour day on average, then the effective hourly wage is $13.64. This is hopefully an underestimate.

10This assumes 7 teachers paid $13.64 per hour score 280 exam in 1.5 hours.

11This assumes 7 teachers eat $30 of pizza.

12Available upon request by contacting Nathan VanHoudnos at nmv@cmu.edu
Figure 5: The output of fictitious T-Study. Turker’s ratings are in the body of the table. Blank cell indicate missing data: student work (row) not rated by that Turker (column).

persistence may fall under the cut-off by chance because not enough information is available to accurately estimate their quality.

In summary, a T-Study is a definition provided by a custom piece of software that allows for ratings to be easily collected from Amazon Mechanical Turk. The output of T-Studies can be thought of as a table, which allows for easy calculation of summary statistics, such as using the rounded average across the Turker-columns as a prediction of the “true” score of the student work.

4.5 Testing the Turkers against the teachers

Here we explore the specifics of our proof-of-concept dataset, describe the T-Studies we rated it with, and then compare the Turkers with the teachers to answer the first research question of the paper: Can anonymous people on the Internet score student work?

4.5.1 The proof-of-concept dataset from Propel EAST

To construct a proof-of-concept dataset we requested from Propel EAST data from their lowest grade level (Grade 3), their highest grade level (Grade 8), and a grade level in between (Grade 6). Propel gave us anonymous photocopies of student exams, a spreadsheet listing the recorded scores from the volunteer teachers keyed to the photocopies by a unique student ID, and the rubrics provided to the volunteer teachers to score the exams.

Table 2 displays summary information about the volunteer teacher scores. The cost column is the approximate amount of money Propel EAST spent to feed the volunteer teachers as a gesture of thanks. It is small. The donation column is a conservative estimate of the value of the time the teachers’ donated. It is comparatively large. The cost and donation columns were derived from the per response estimates calculated at the end of Section 4.4.1. The question column gives the number of the hand-scored constructed response items. Note that since each grade has only two, the proof-of-concept dataset is small. This limits our ability to generalize our results.

<table>
<thead>
<tr>
<th>Grade</th>
<th>No. Students</th>
<th>Administration</th>
<th>Cost*</th>
<th>Donation*</th>
<th>Question</th>
<th>Ave. Score</th>
<th>Standard Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>38</td>
<td>October 2010</td>
<td>$8.14</td>
<td>$38.86</td>
<td>Question 35</td>
<td>2.21</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Question 36</td>
<td>2.97</td>
<td>0.94</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
<td>October 2010</td>
<td>$10.50</td>
<td>$50.11</td>
<td>Question 29</td>
<td>1.24</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Question 30</td>
<td>2.37</td>
<td>1.01</td>
</tr>
<tr>
<td>8</td>
<td>43</td>
<td>March 2010</td>
<td>$9.22</td>
<td>$43.98</td>
<td>Question 26</td>
<td>1.35</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Question 27</td>
<td>0.86</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Table 2: Summary of the student scores from the pizza party results provided to us by Propel EAST. *Approximate, for derivation see Section 4.4.1.

Copies of the rubrics can be found in Appendix A. We give a summary of the remaining rubrics in Table 3. Four of the rubrics are straightforward to split into scaffolded judgments for the Turkers. We refer to these rubrics as additive rubrics because the subscores of the judgments can be added together to get the
total score. Two of the rubrics combine the evaluation of multiple criteria into a single judgment, so the interface for the Turkers must manage this additional complexity. We refer to these rubrics as *simultaneous* rubrics because the volunteer teacher must simultaneously determine multiple criteria to answer a judgment.

<table>
<thead>
<tr>
<th>Additive</th>
<th>Simultaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3 Question 35</td>
<td>Grade 3 Question 36</td>
</tr>
<tr>
<td>Grade 6 Question 29</td>
<td>Grade 6 Question 30</td>
</tr>
<tr>
<td>Grade 8 Question 26</td>
<td>Grade 8 Question 27</td>
</tr>
</tbody>
</table>

Table 3: Overview of rubric types.

The first simultaneous rubric is Grade 6 Question 30 which we discussed in Section 4.3.2. The second is Grade 3 Question 36. Figure 6 excerpts the relevant portion of the rubric. It asks for two criteria. The first is “Rate the quality of the definition of symmetry.” and the second is “Rate the quality of the explanation.”. Both criteria are implicitly rated on a scale of Correct/Partial/Incorrect, but the judgment asked by the rubric combines the two. It specifies that the student is to receive 1.5 points if both criteria are correct (Correct/Correct), 1 point if only one is correct (Correct/Incorrect or Incorrect/Correct), and 0.5 if either is partially correct (Partial/Incorrect or Incorrect/Partial). Note first, that this is not an additive scale: There is no value of the two criteria such that the criteria can be scored separately and then added (e.g. Correct = 0.5, Incorrect = 0, fails at Correct/Correct). Note second, that the rubric does not specify explicitly the score for the following combinations: Correct/Partial or Partial/Correct. That is left to the discretion of the expert teachers.

![Figure 6: 4Sight Rubric Grade 3 Question 36 Part B](image)

Figure 6: 4Sight Rubric Grade 3 Question 36 Part B

It is not possible to determine if the inter-rater reliability of the volunteer teachers varied according to rubric type. This cannot be calculated after the fact because the design of the pizza party did not compare a subset of students between teachers or identify which teacher scored which student. Since we did not conduct a reliability study during a pizza party, we can only calibrate the expected reliability from the literature. Due to the semi-informal one time nature of the pizza-party process, we choose the rubric literature as a point of comparison. Specifically, we rely on a recent review by Jonsson and Svingby (2007) when we compare the Turkers with the teachers at the end of this section.
4.5.2 Rating each exam with a size 5 and a size 30 T-Study

We rated each grade level in the proof-of-concept dataset with two T-Studies. Tables 4 and 5 summarize the T-Studies. The first table depicts the total number of scores collected from the Turkers and the cost to collect those scores. The Grade 8 T-Studies have roughly twice as many scores because each of the two parts of each question had its own HIT (i.e. Question 26 Part A, Question 26 Part B, Question 27 Part A, and Question 27 Part B). This was due to the length of the question parts. We assumed that since each part took a whole page, which was roughly the same amount of space as a whole question from the lower grades, then splitting the questions into two HITS was necessary. Since we pay the Turkers $0.03 per score and Amazon Mechanical Turk a flat fee of $0.005 per score collected from the Turkers, the total cost of the entire T-Study is easily calculated.

![Table 4: The total number of scores collected from Amazon Mechanical Turk and the associated cost to collect those scores. Note that Grade 8 has 2 HITS per question because each part of the Grade 8 questions was equivalent to an entire question from the lower grades.](image)

The second table, Table 5 describes the characteristics of the Turkers who participated in the T-Studies. Overall the T-Studies took a fairly long time to complete; the size 5 T-Studies took hours and the size 30 T-Studies took tens of hours. For the Turkers, the table reports medians instead of means because of the skewed distributions of the reported statistics. In all cases, the median Turker did not rate very many rows. Grade 8 stands out because the Turkers spent one-fourth to one-half of the time working on an individual HIT as compared to the lower grades. This suggests that splitting the Grade 8 questions into two HITs was unnecessary, and therefore unnecessarily expensive.

![Table 5: Summary table of T-Studies collected on Amazon Mechanical Turk](image)

Figure 7 explores the persistence of the Turkers in greater depth. It compares the persistence distributions between the T-Studies by normalizing them with the number of rows in a given study. The smaller size 5 T-Studies are on the left and the larger size 30 T-Studies are on the right. Relative to the amount of work available to a given Turker, the distribution of how many rows any given Turker will choose to score did not change by either T-Study size or grade level. Pair-wise testing for differences in distribution using a modified Kolmogorov-Smirnov test\(^{13}\) and applying the Bonferroni correction for multiple testing found that only the Grade 3 size 30 T-Study was significantly different at the 5% level. It differed from both the Grade 8 T-Studies and the larger, size 30 Grade 6 T-Study.

We hypothesize that this difference is detected because the level of the mathematics in Grade 3 is much lower than for the other grades and more Turkers likely felt more confident about grading this work. However, it is striking that the Grade 8 T-Studies were not found to be significantly different given that the Turkers who participated in them had higher wages. This suggests that although the Turkers from Grade 8 were paid more, they did not attempt to do more work. This is unexpected as the literature suggests that paying the Turkers more raises their persistence (Mason & Watts, 2009; Shaw et al., 2011).

Figure 7: Distribution of persistence normalized between T-Studies.

4.5.3 Comparing the Turker scores with the teacher scores

We first compare the Turkers and the teachers by calculating the aggregate score of the T-Study. We use the rounded average of all the Turkers who rated a given row. We then count the number of times the aggregate T-Study score mis-classifies a particular piece of student work as having rating X when it should have rating Y. Collecting those mis-classifications into a table yields the confusion matrices that Figure 8 depicts. If the aggregate T-Study score correctly classified every single piece of student work, the only non-zero entries in the confusion matrix would be on the diagonal. If the Turkers in the T-Study were more lenient than the teachers, the non-zero entries would primarily be in the bottom right of the confusion matrix. Similarly, if the Turkers were harsher than the Teachers, the non-zero entries would primarily be in the top left of the confusion matrix. If the Turkers are unbiased but unreliable, the non-zero entries would spread out evenly from the diagonal instead of concentrating only on those diagonal cells. The figure has been shaded so that darker cells represent where most the mass falls in a given row. This simplifies visual comparison between the groups of students who received the same teacher score. The blue represents Grade 3, green represents Grade 6, and red represents Grade 8. This is a coloring convention that we use throughout the paper. In addition, the diagonal cells are outlined in black to guide the eye.

The Turkers are fairly close to the teachers in all cases. The majority of the mass in each of the confusion matrices is either on the diagonal or very near to it. In addition, the Turkers do not have a systematic bias. In Grade 8 the Turkers are harsher; in Grade 6 the Turkers are more lenient.

Although we would expect the mass to be more concentrated on the diagonal for the size 30 T-Studies, the two T-Study sizes look very similar. It is not immediately clear if the size 30 T-Studies do any better than the size 5 T-Studies at predicting the teacher’s scores. To quickly compare the confusion matrices with each other and with the literature, we calculate Pearson’s correlation coefficient. We chose this statistic because it is interpretable and commonly used.

Of the twelve correlations reported in Figure 8, eight are bolded because they are above 0.70, which is a common reliability threshold in the literature (Jonsson & Svingby, 2007). Note that with the exception of the simultaneous Grade 3 rubric (i.e. Question 36) all of the size 30 T-Studies had correlations above
### Teachers versus Turkers

#### Confusion Matrices

<table>
<thead>
<tr>
<th>Grade 3</th>
<th>Question 35</th>
<th>Additive</th>
<th>Size 5</th>
<th>Size 30</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Size 5   Size 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.95      .95</td>
</tr>
</tbody>
</table>

| Question 36 | Simultaneous | Size 5 | Size 30 | Pearson Correlation | .64 | .66 |

<table>
<thead>
<tr>
<th>Grade 6</th>
<th>Question 29</th>
<th>Additive</th>
<th>Size 5</th>
<th>Size 30</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Size 5   Size 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.81      .85</td>
</tr>
</tbody>
</table>

| Question 30 | Simultaneous | Size 5 | Size 30 | Pearson Correlation | .69 | .75 |

<table>
<thead>
<tr>
<th>Grade 8</th>
<th>Question 26</th>
<th>Additive</th>
<th>Size 5</th>
<th>Size 30</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Size 5   Size 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.65      .75</td>
</tr>
</tbody>
</table>

| Question 27 | Additive | Size 5 | Size 30 | Pearson Correlation | .85 | .79 |

---

Figure 8: Comparing the rounded average of the Turker scores with the volunteer teacher scores. The confusion matrices are shaded by row-weight.
the 0.70 threshold while only half of the size 5 T-Studies met that threshold. This suggests that increasing T-Study size can improve the quality of the aggregate T-Study score.

More importantly, this answers the question of the section: Can anonymous people on the Internet score student work? The proof-of-concept data suggests that yes, they can. The next section expands on this by determining the size of the T-Study necessary to consistently achieve acceptable results.
5 How many of these anonymous people must score a student response before the average of their scores agrees with a teacher?

This section asks what size of T-Study is required before the correlation between the aggregate T-Study score and the teacher score is significantly higher than a specified lower bound. Our approach is to develop a cost-effective method to predict Turker performance as function of T-Study size.

Although the Turkers are cheap, they are not free. Adding the size 5 costs reported in Table 4 and dividing by 5 shows that it costs $12.11 to score the entire dataset once at size 1. This does not seem like much, but costs quickly increase if we study the performance of the average Turker score as a function of T-Study size. If we pick our T-Study sizes as 3, 5, 7, 10, and 15, then it would cost $484.40 to rate the dataset once.

To answer the guiding question of this section requires a further increase in costs because we must repeat the T-Study enough to determine confidence intervals for the correlation coefficient between the Turkers and the teachers. To make a conservative estimate of the number of these repetitions required, we make a series of simplifying assumptions about the values reported in Figure 8. First, we assume that the correlations are affected by the T-Study size and not by grade level or question type. This allows us to pool them together to estimate the variability of the rating process. We further assume that the size 5 correlations are independently and identically normally distributed with an unknown mean and variance. These assumptions allow us to match the variance of the sample mean and work backwards to the variance of the assumed distribution. This in turn allows us to calculate that a 95% confidence interval with an error of no more than 0.20 units of correlation would require the size 5 T-Study be repeated at least 10 times. This is a conservative estimate: If the underlying distribution is either skewed or has thicker tails than a normal distribution, then more samples would need to be collected.

If we further assume that each T-Study size would need the same number of repetitions as size 5 T-Study, then the total cost would be $4,844. This is unnecessarily expensive because it ignores that a large T-Study effectively contains smaller T-Studies within it. In this section, we present a method that achieves the same results, but costs an order of magnitude less, specifically $423.80.

To clearly discuss the previous literature relevant to our new method, we must define a more general term for Assignments on AMT whose output can be thought of as a T-Study table with the “responses” on the rows and the Turkers on the columns. We refer to these kinds of AMT Assessments as Experiments.

Through this section we use the term Experiment for general application and the term T-Study when referring to specific examples drawn from our work. We begin with a discussion of how others have extracted many smaller sized Experiments from a single larger sized Experiment. We then more closely examine the process of how Turkers interact with Experiments to propose a new time-based method of subsampling a large size Experiment into a series of smaller size Experiments. We then apply this method to the T-Studies we collected to calculate confidence intervals for the correlations between the Turkers and the teachers.

5.1 Previous work on subsampling a large size AMT Experiment

Previous work on subsampling large size Experiments focused on the output of the Experiment by using the table metaphor shown in Figure 5 on page 15. Snow, O’Connor, Jurafsky, and Ng (2008) was the first to use this metaphor to create smaller Experiments by subsampling from a single large Experiment. Their approach to creating a subsampled Experiment of size \( S' \) is to select \( S' \) scores at random from each row. This treats the cells of the table as independently and identically distributed (iid).

This iid assumption is too strong because it does not allow for the construction of models that correct for between-Turker or within-Turker variability. For example, any model that attempts to correct for between-Turker variability will be sensitive to the Turker’s persistence, the number of rows a Turker has completed. To accurately predict the performance of a given model at size \( S' \), we want the persistence of the average Turker in a subsampled Experiment to be equal to the average persistence of an independently collected Experiment of size \( S' \). If the sub-sampled average persistence is less than the observed average persistence, then the worker sample size is not as large as it should be, and any model that accounts for between-Turker variability will have fewer degrees of freedom to make estimates. This will cause predictions of model performance to be wrong. Specifically, low-persistence subsampling will preserve less data and will underestimate the predicted performance on a newly collected Experiment of the same size.
Since Snow et al.’s method randomly selects Turker scores for a given row in the scoring table (Figure 5), an individual Turker has an $S'\over S$ chance at being selected if they provided a score for that row. For a Turker that scored $p$ rows of the larger Experiment, then the number of rows they will “score” in a sub-sampled Experiment ($p'$) is binomially distributed as they have $p$ equal chances for being successfully included with probability $S'\over S$. This implies that their expected persistence over many such sub-samples is

$$E[p'] = S' \times p \quad \text{(1)}$$

This result would imply that the average of the persistence of the Turkers from size 5 Experiments should be one-sixth of the size 30 Experiments. This is not the case. Calculating the means instead of the medians for Table 5, finds that the average Turker persistence for the size 30 T-Studies ranges from 13 rows to 40 rows, and the size 5 average persistence ranges from 13 to 29 rows. This is not near one-sixth.

Although it is possible to preserve persistence by relaxing the assumption of iid cells throughout the table to an assumption of iid cells conditional on a particular Turker’s column, this kind of sampling-by-column cannot address the time-series nature of within-Turker variability. Larger experiments take longer to complete. This gives Turkers more opportunities to score the work over several sessions. For example, close examination of the Grade 6 size 30 T-Study reveals several Turkers who return the next day to rate student work. This cannot occur in the size 5 T-Studies because those T-Studies last only a few hours. To the extent that scores from a given Turker vary from day to day (e.g. they learn), equally including the scores from the second day with the scores from the first day will introduce a bias. This bias will affect even the simple estimate of averaging the Turker’s scores.

### 5.2 A new idea for predicting quality as a function of Experiment size

Instead of assuming iid cells or iid cells conditional on a given column, we deemphasize the metaphor of the output table. We focus instead on the process that generated the Experiment to develop a new subsampling method that accounts for both between-Turker and within-Turker variability.

We bootstrap the generating process of the Experiment to account for the between-Turker variability and use a time-based subsampling procedure to account for the within-Turker variability. By combining these two steps we can generate an arbitrary number of “pseudo-experiments” across a range of experimental sizes. These pseudo-experiments are bootstrap replications that approximate the underlying generating process, so we use them to both derive point-estimates as a function of experimental size as well as the associated confidence intervals. This allows us to predict the smallest T-Study size needed to ensure that the correlation between the Turkers and the teachers achieves a minimum quality threshold.

#### 5.2.1 General assumptions about AMT Experiments

Experiments unfold in two phases. The first, the beginning of the Experiment is when every row has at least one more rating to achieve before reaching the requested size. The second phase, the endgame of the Experiment is when at least one of the rows has reached its required size.

In beginning of the Experiment, every new Turker sees all of the rows as available for scoring. Each row the Turker scores decreases the number of rows available to him or her because the Turker may only score a row once. In the endgame of the Experiment, the number of rows a given Turker sees as available becomes a function of both the work that the particular Turker has done and the work that other Turkers are doing in parallel. Other Turkers can finish a row while the particular Turker is still working on a different row. This causes the number of available rows to decrease at a faster rate. The change of this rate will be more pronounced for smaller Experiments and less pronounced for larger Experiments. Accordingly, it is the negative acceleration of the number of available rows that gives a Turker his or her only indication of experimental size. Therefore its is unlikely that experimental size affects Turker behavior.

Since experimental size should not affect the behavior of Turkers, it is reasonable to create smaller Experiments by selecting a subset of the Turkers from the larger Experiment. The choice of subsampling procedure requires careful thought as the Turkers have both within-Turker and between-Turker variability. The within-Turker variability is assumed to be the result of the passage of time. For example, Turkers may
Figure 9: Turker quality versus entry time. Grade 8 is missing because the Turkers scored the student work at the part-level instead of the question-level for those T-Studies.

learn as they score more student work, or Turkers may get bored towards the end of their scoring sessions. Therefore, preserving the time-series nature of an individual Turker’s ratings will lower the subsampling bias.

We assume that the time the Turker first enters our task is dependent upon factors such as the time-of-day, the availability of other tasks on the AMT Marketplace, and the characteristics of the task itself, but that the entry time is independent of Turker quality. This is a reasonable assumption because Amazon Mechanical Turk is a world-wide marketplace where it is always someone’s prime local time for working. We directly check this assumption by calculating the correlation coefficient between each Turker and the teachers’ scores. Figure 9 displays a scatter plot of these correlations against the entry time of the Turkers for the Grade 3 and Grade 6 size 30 T-Studies (Grade 8 is missing because the Turkers rated the questions at the part-level while the teachers rated at the question level). The figure shows no evident association between entry time and Turker quality. This allows us to construct a bootstrap procedure that preserves the rating order within a Turker but mixes the rating order between Turkers.

5.2.2 Time based subsampling of a large experiment

The proposed time-based subsampling procedure considers the Turker scores in the order that they were received and places them into the scoring table as long as the appropriate row has not yet reached the desired size. The rest of the scores are simply discarded. Here we explain the consequences of discarding those scores.

To clearly discuss these consequences, we define the row-size as the number of Turkers who have completed a given row. Row-size varies over time as more Turkers do more work. For example, Figure 10 plots the average row-size, the maximum row-size, and the minimum row-size as a function of time for the larger size 30 T-Studies. To extract a size 10 T-Study all of the rows must have at least 10 scores. As can be seen from the graph, at any given time there is wide variation in the distribution of row-sizes. For example, when the minimum row-size is 10, the dotted vertical line intersects the average row-size between 16 scores and 20 scores for these three T-Studies. Therefore restricting one of these T-Studies to a size of 10 would remove between 6 and 10 scores per row. Given that the median Turker scores about that many rows (Table 5), that is a lot of data to remove.

We refer to a score that the sub-sampling procedure removes as a censored score. To explore how censoring affects individual Turkers, we subsample the size 30 Grade 3 T-Study to a size of 10 and depict the results in Figure 11 with a simplified version of Downs et al.’s (2010) time-line view for AMT Experiments.
Figure 10: Average, minimum, and maximum row-sizes as a function of time for the size 30 T-Studies. The values of the various row-sizes are reported when the minimum row-size reaches 10. The endgame of the T-Study is also highlighted.

The colored dots represent scores that are included in the subsample; the black dots censored scores. As the inset figure shows, in the start of the endgame, black dots are rare; towards the end, the colored dots are rare. For Turkers that span a wide range of time, it is possible that the subsample will select scores in the beginning of the Turker’s employment, ignore scores in the middle, and select additional scores at the end. This is precisely what the main figure displays for Turker 48.

If Turker 48 exhibits within-Turker variability, the inclusion of his or her last scores while censoring the middle scores will bias the subsample. This out-of-order inclusion in the T-Study can occur only during the endgame. Since the endgame lasts a limited amount of time, this time-based method is expected to have lower bias compared to the iid methods which always include Turkers out-of-order. The process of censoring in this time-based method mirrors the process of rows disappearing from an individual Turker’s queue of available work that occurs during the endgame of a collected experiment.

Figure 12 displays a comparison between subsampling the size 30 T-Studies to a size of 5 with the collected size 5 T-studies. The left column displays the observed T-Studies; the right column displays the sub-sampled T-Studies. The two sets of T-Studies do not look very similar. The collected T-Studies score the same amount of student work in less time (i.e. the plot doesn’t extend as far on the time axis ) with fewer Turkers ( i.e. the plot is not as tall on the number of Turkers axis ). Since we assume that the entry times of the Turkers do not affect their quality, we can dismiss the variation on the time-axis as irrelevant. However, the variation in the number of Turkers must be addressed.

The number of Turkers in an Experiment is inversely proportional to the average persistence of all the Turkers in the Experiment. That the collected T-Studies used fewer Turkers to complete the same work implies that the individual Turkers in the collected T-Studies each did more work than the corresponding Turkers from the subsampled T-Studies, their persistence was lower. Since a goal of the subsampling approach is to preserve persistence, we must account for this apparent downward bias to the persistence estimate. This
Figure 11: Visualization of time-based subsampling’s endgame construction. The colored dots represent scores that are included; the black dots represent scores that are not. The inset figure displays the entire T-Study while the main figure displays a detail of the endgame.
motivates the remainder of the section: Is this apparent downward bias significant? We address this question by estimating the variability of persistence with the bootstrap.

5.3 Bootstrapping a large experiment

We explore the bootstrap as a method to estimate the variability of the scoring process to quantitatively determine if the sub-sampling procedure matches the collected experiments. This will complete our method to predict the performance of Turkers as a function of T-Study size.

Bootstrapping these data requires careful thought. Our goal is to capture the expected variability of drawing a new set of Turkers to rate the same student work. This is a different quantity from the expected variability of taking a new set of student work and giving it to the same Turkers. To continue with the table metaphor, we wish to estimate the variability of the Turker-columns, not the variability of the student-rows. Therefore, we cannot average across the columns to calculate the output of the Experiment and then create bootstrap replications by sampling from that output vector. That procedure estimates the student-row variability by averaging out the variability of the Turkers. To calculate the variability of the Turker-columns we must develop a bootstrap procedure based on the process of the Experiment. In that way, we do not average out the Turker-columns, but instead preserve between-Turker variability.

Since the entry times of the Turkers are independent of Turker quality, we mix the Turkers with each other’s entry times. This allows for Turkers that first appear during the end of the large Experiment to still participate in the beginning of a small bootstrapped Experiment. The specific procedure is as follows:

1. Draw from the set of entry times ($E$) with replacement to create a new set of equal length ($E^*$).
2. Draw from the set of Turkers ($T$) with replacement to create a new set of equal length ($T^*$).
3. Match the first Turker in set $T^*$ with the first entry time in set $E^*$ and then shift all of that Turker’s work to line up with the new entry time.
4. Repeat step 3 for all of the ($T^*, E^*$) pairs.

Figure 13 illustrates the output of this process. The generating T-Study is in color in the upper left-hand corner. The bootstrap replications are plotted in black. The bootstrap replications look similar to each other and to the generating T-Study. This is expected. However, the bootstrap replications do not have an endgame: they have no size restriction to force the censoring of ratings towards the end of the T-Study.

5.4 Combining bootstrapping and subsampling

Figure 14 explores this lack of censoring by size more directly using the three size 30 T-Studies as examples. The darker lines are the minimum and maximum row-sizes of the generating T-Study and the lighter lines are the bootstrap replications of that T-Study. The dotted vertical line marks the transition to the endgame of the generating T-Study. In all cases, in the beginning of the T-Study, the bootstrap replications track the generating T-Study closely. During the endgame the censoring of the generating T-Study causes it to diverge from the bootstrap replications.

During the endgame, censoring causes the minimum row size curve to bend upward until it matches the desired size of the experiment. Since the bootstrap replications lack this upward pressure, their minimum row size curves stabilize instead of bending upward. This limits the ability of a bootstrap replication to predict beyond the size that the minimum row-size curve achieves. For example, if the minimum size curve ends at 16 scores per row, then that replication can make no predictions for a size 17 experiment since the replication is missing at least one rating.

This makes it unwise to interpret the results of subsampling a bootstrap replication beyond the experimental size at which the minimum row-size curve begins to typically flatten. A convenient method to find the size where this occurs is to look at the success rate for subsampling a bootstrap replication for a given size.

In Table 6 we report the number of successes for subsampling 100 bootstrap replications to the requested size. Since we will use these replications to determine 95% confidence intervals, we set the size exclusion threshold at a 5% failure rate. Therefore we can make predictions with the subsampled replications at sizes
Figure 12: A visual comparison of a time-based sub-sample of the size 30 T-Studies with their independently observed size 5 counterparts. The observed T-Studies are on the left; the sub-sampled T-Studies are on the right.
Figure 13: A visual comparison of the bootstrap replications (black) with the Grade 6 size 30 T-Study used to create the replications (green).
Figure 14: Bootstrap replications of minimum and maximum row size. The region of the graphs to the right of the dotted line represents the endgame of the generating T-Study.
Table 6: Number of successes when sub-sampling 100 bootstrap replications to the specified size.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Size 3</th>
<th>Size 5</th>
<th>Size 7</th>
<th>Size 10</th>
<th>Size 15</th>
<th>Size 20</th>
<th>Size 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>71</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>97</td>
<td>79</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>75</td>
<td>33</td>
</tr>
</tbody>
</table>

lower than 15, but not for any larger sizes. At the larger sizes, the generating size 30 T-Studies cannot supply enough Turkers to approximate an independently collected T-Study.

5.5 Comparing to previous work

We define a *Pseudo-Experiment* as a successful subsampling of a bootstrap replication. Since these Pseudo-Experiments are themselves bootstrap replications, they can be used to estimate confidence intervals for arbitrary statistics by calculating the statistic across the set of Pseudo-Experiments and taking the appropriate quantiles of the resulting set.

The estimation of confidence intervals allows us to compare the average persistence of the Turkers in the Pseudo-Experiments against the expected persistence of Snow et al.’s iid subsampling approach. Figure 15 visualizes this comparison. The horizontal thick line is the average persistence of the Turkers in the size 30 T-Study. The diagonal dashed line is the expected persistence of the iid subsampling procedure calculated from Equation 1. The solid dots connected by a solid black line represent our prediction of persistence as a function of size; they are the mean of the average persistence calculated across the Pseudo-Experiments. The gray line and open circles also represent the mean of the average Pseudo-Experiment persistence, but given that more than 5% of the Pseudo-Experiments failed at a size of 20 or greater, we do not give them as much credence. The polygons surrounding the point estimates represent 95% confidence intervals of average persistence. The gray square is the persistence of the independently collected size 5 T-Study.

The confidence intervals do not overlap the expected persistence from Snow et al.’s iid method. Therefore, the Pseudo-Experiment method significantly preserves persistence compared to prior work. The gray squares fall within the uncorrected 95% confidence interval in two out of three cases. Correcting the confidence interval for three simultaneous tests with the Bonferroni correction widens the confidence intervals to include all of the independently collected points. This suggests that the Pseudo-Experiment method may accurately predict persistence for independently collected T-Studies. If the Bonferroni correction removes our power to detect a difference between the collected T-Studies and the Pseudo-Experiments, then the new Pseudo-Experiment method is at least a tighter bound on predicted persistence compared to Snow et al.

In summary, our new method improves upon the existing approach because it accounts for the nature of the task instead of merely considering the output of the task. Our method is therefore able to address within-Turker and between-Turker variability. In the next subsection, we apply this method to deriving confidence intervals for the correlation between the Turkers and the teachers to answer the second research question of the paper: How many of these anonymous people must score a student response before the average of their scores agrees with a teacher?

5.6 Comparing the Turker scores with teacher scores for a range of T-Study sizes

Recall that Figure 8 in Section 4.5.3 reported single values of the Pearson correlation coefficient for comparisons of the aggregate T-Study score and the rating from the volunteer teachers. Now that we developed a method to find confidence intervals for the correlation coefficients instead of merely point estimates, we can choose the lowest level of reliability that is acceptable. After fixing this quality threshold, we can determine the T-Study size that predicts when we will exceed the threshold with 95% confidence.

We rely on Jonsson and Svingby’s (2007) review of the rubric literature to determine an acceptable threshold for the Pearson correlation coefficient. The review reports the range of the “majority of estimates” of the Pearson correlation coefficient as $\rho = 0.55$ to $\rho = 0.75$. We therefore set our threshold of Turker quality at $\rho = 0.55$. This ensures that the scores the Turkers provide match the teachers by at least achieving a

30
Figure 15: Persistence of Pseudo-Experiments compared with iid sub-sampling of Snow et al. (2007) and the independently collected size 5 T-Study.
reliability marginally accepted by the literature. Therefore, when the Pseudo-Experiment confidence interval crosses $\rho = 0.55$, we take that experimental size as the minimum acceptable size.

Figure 16 displays the correlation between the aggregate T-Study score, which is the rounded average Turker score, and the score the teachers assigned for a range of sizes. The two colored squares at sizes 5 and 30 represent the values from the independently collected T-Studies that are reported alongside the confusion matrices in Figure 8. The solid dots connected by a solid black line represent the typical correlation predicted by the Pseudo-Experiment method. The extension of this prediction to hollow circles and a gray line represents the inability to construct a confidence interval due to too many Pseudo-Experiment failures at those large sizes. The colored polygons represent the uncorrected 95% confidence intervals obtained from calculating the quantiles across the set of Pseduo-Experiments. The horizontal lines at $\rho = 0.55$ and $\rho = 0.75$ represent the ranges from the 2007 review. This divides the graph into “Better”, “Okay”, and “Worse” sections.

There are three things to note about this figure. First, consider only the leftmost column. These three graphs represent the combined performance of the average of Turker scores across both questions in a given grade. The median performance is rather flat; there is not a practical improvement in the expected performance between size 3 and size 15. What does change is the variability. The variability at sizes 3 and 5 is quite large when compared to the variability at size 15. This suggests that the primary effect of adding more Turker scores is to reduce the variability of the average estimate, not improve its median-performance. By size 10, the predicted confusion matrices are no longer different from the size 30 confusion matrices that were depicted in Figure 8. The performance also varies by grade; the confidence intervals between the Grade 3 Size 10 Pseudo-Experiments and the Grade 8 Size 10 Pseudo-Experiments do not overlap. They are both practically, since the typical Grade 3 correlation falls in the “Better” range while the Grade 8 is merely “Okay”, and significantly different at a size of 10.

Second, consider the differences between the additive and simultaneous rubrics. It is clear the difference in the quality of the average Turker score is both practically and significantly different between the additive and simultaneous rubrics for Grade 3. The differences for Grade 6 are less pronounced, but there appears to be a practical difference even if it is only marginally significant at size 15. In addition, as more Turkers are added, the variability of the simultaneous rubrics decreases more slowly than the variability of the additive rubrics. Both of these effects are more pronounced for Grade 3 than for Grade 6. These two examples are not enough to generalize from, but we speculate that the effects could be from the additional cognitive load the teachers faced when grading with a simultaneous rubric. Particularly, since the Grade 3 Question 36 rubric intentionally left part of the scoring decision to discretion of the teachers, the implementation of the rubric for the Turkers that removed this decision and correspondingly decreased the cognitive load of the task could have biased the average Turker rating away from the teachers. This suggests that the teachers and Turkers are more likely to agree when the rubrics they use require similar amounts of mental effort and discretion.

Finally, the figure gives the minimum acceptable size for a T-Study. In all cases, the additive rubrics cross the threshold between “Worse” and “Okay” at least by size 7. The simultaneous rubrics do not cross this threshold until size 10. We therefore choose 10 as the answer to the question posed by this section: How many of these anonymous people must score a student response before the average of their scores agrees with a teacher?

---

See Section 4.5.1 for a brief discussion.
Figure 16: The Pearson correlation between the rounded average Turker score and the score the teacher assigned.
6 Discussion

This paper sought to solve a specific need of a local charter school in a proof-of-concept demonstration that anonymous people from a commercial crowdsourcing service could cheaply and reliably score student work. The hope was that the charter school could solve their problem with this new method, but also that the new method could more generally reduce the cost in teacher-time of scoring constructed response assessments.

The main result of the paper is that the correlation between the average of 10 Turker scores and the scores a teacher assigned had a typical value between $\rho = 0.86 \pm 0.04$ for the Grade 3 exam to $\rho = 0.72 \pm 0.07$ for the Grade 8 exam and is predicted to be at least $\rho = 0.55$ with 95% confidence for every question in every exam. However due to the limited sample size of the proof-of-concept dataset, these results may not generalize beyond the rating of constructed response mathematics questions from exams similar to 4Sight. In addition, the inter-rater reliability of the gold-standard dataset we use for comparison is unknown. It is possible that unnecessary noise in the teacher ratings artificially lowers the performance of the Turkers. It is also possible that the Turkers are biased away from the teacher’s performance by the design of the rubric meant to scaffold them.

The cost of our new method falls between the cost Propel incurred to offer pizza to the volunteer teachers and the value of the time the teachers donated to the scoring of student work. Propel paid about $0.11 per student response and the teachers donated at least $0.51 of their time per student response, while the cost of crowdsourcing the student work with 10 Turkers was $0.35 per student response. That the Turker’s cost more than Propel is currently spending is not surprising. The Turkers are not volunteering. The Turkers, however, are sufficiently affordable that it may be reasonable to free teacher time for other, higher value tasks such as reflecting, either alone or in a group setting, on the responses the students gave without being burdened by the additional cognitive load of scoring the response at the same time.

A secondary result of the paper is the development of a novel method to subsample a large Experiment on Amazon Mechanical Turk into a series of smaller Pseudo-Experiments of varying size. These Pseudo-Experiments can calculate confidence intervals for arbitrary statistics as a function of size. This new method costs much less, in our case a factor of 10 less, than directly collecting those Experiments. The new method improves on previous work by relaxing the assumptions on Turkers so that both within-Turker and between-Turker variability can be accounted for.

Future work could model the Turkers so that the ratings of ‘bad’ workers could be down-weighted or removed while the ratings of ‘good’ workers were up-weighted in the final estimate. It is clear from the marginal distribution of the correlations computed in Figure 9 that our use of the sample mean in this paper decreased the eventual agreement with the gold-standard ratings because of few Turker’s who were of very low quality moved the estimate farther away from the teacher’s rating.

It was, however, difficult to find those offending Turkers without using the gold-standard scores. Our preliminary work on modeling the Turkers found no models that outperform the rounded mean. The primary difficulty is that the within-Turker sample size is very low. We conjecture that the majority of the Turkers do not rate enough student work for the model to learn about them. This renders common models from the crowdsourcing literature such as those based off of Dawid and Skene (1979) as inappropriate because many Turkers do not provide enough ratings to match the number of parameters in the model. We have attempted to fit simple OLS regression models that account for Turker bias and GLS models that account for Turker variance (Wilson, 1988), but these models yield estimates that are neither significantly nor practically different from the simple mean. We have also attempted partially-pooling the Turkers in an empirical Bayes framework with a simple normal model (Efron & Morris, 1975), but this too did not succeed. Future research could consider hierarchical approaches such as followed by Patz, Junker, Johnson, and Mariano (2002).

Another direction for future research is exploring the feasibility of crowdsourcing the scoring of student work beyond constructed response mathematics questions. Other subjects and other assessments may require different kinds of rubrics, which may or may not be suitable for translating into interfaces for Turkers. Of particular interest for future research are the assessments teachers develop themselves, since reducing the scoring burden of any assessment that teachers wish to use would have the highest expected utility.

The hope of this research is that it would give teachers more time to focus on the kinds of activities they enjoy most, such as figuring out what students are thinking and how to shepherd them to higher levels of understanding. It is with this in mind that we present this proof-of-concept to the research community for further refinement and improvement.
References


A 4Sight Rubrics

A.1 Grade 3 rubrics

Problem 35 Figure 17 depicts the first Grade 3 constructed response question. It deals with counting, comparing, and making change when presented with graphical representations of US money. The rubric asks the teachers to make a series of four Yes/No judgments. Each Yes is worth one point.

Figure 17: 4Sight Rubric Grade 3 Question 35

Problem 36 Figure 18 depicts the second Grade 3 constructed response question. It asks students to demonstrate their understanding of symmetry. Although Parts A and C ask for simple Yes/No additive judgments, Part B is requires two criteria to judged simultaneously. For example, the presence of both a correct definition and correct explanation earns 1.5 points, but the presence of only one of either definition or explanation earns 1 point (instead of 0.75 points).
(Test 2)

A. Look at the following figures. **Draw** a circle around the figure that does **not** show a line of symmetry.

B. **Explain** how you know the line in this figure is **not** a line of symmetry.

C. Look at the shapes below.

Find the figure that is symmetrical. **Draw** one line of symmetry on the figure.

(36)

(Test 2) Top Scoring Response

Part A

Answer:

Pentagon (Students should circle the pentagon.)

(1 score point)

1.0 point for correct answer

Part B

Explanation:

A line of symmetry divides a shape into two **equal** parts so that one part can fold over and lie exactly on top of the other. The line on the pentagon is **not** a line of symmetry because the two parts of the pentagon are **not** the same size.

(1.5 score points)

1.5 points for correct definition of a line of symmetry and explanation of why the line drawn on the pentagon is **not** a line of symmetry or

1.0 point for correctly explaining why the line on the pentagon is **not** a line of symmetry without defining a line of symmetry or

1.0 point for correctly defining a line of symmetry without explaining why the line on the pentagon is **not** a line of symmetry or

0.5 point for not enough information to completely explain what a line of symmetry is or why the line on the pentagon is **not** a line of symmetry

Part C

Answer:

(1.5 score points)

0.5 point for drawing a line on the symmetrical figure and

1.0 point for drawing a correct line of symmetry

NOTE: This shows only one of the possible lines of symmetry. All reasonable responses should be accepted. Student answers will vary.

Figure 18: 4Sight Rubric Grade 3 Question 36
A.2 Grade 6 rubrics

Problem 29 Figure A.2 depicts the first Grade 6 constructed response question. It deals with reasoning about fractions. Although the rubric evenly splits the four possible points between Part A and Part B, each of the two Parts has a different allocation of its two points between the Yes/No judgment of "Is this correct?" and the more nuanced judgment of "What is the quality of the explanation?". Part A awards 1 point for the correct answer and then asks the rater to judge the explanation on a scale with three levels (1 point, 0.5 points, and 0 points). In contrast, Part B places more emphasis on the explanation by awarding only 0.5 points for a correct explanation and adding an additional (1.5 points) level to scale from Part A.

29. **(Test 2) Top Scoring Response**

<table>
<thead>
<tr>
<th>Part A</th>
<th>Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answer:</strong> Shirley made 40 pieces of ribbon.</td>
<td><strong>Work/Explanation:</strong> I divided the length of ribbon by the fraction to find the number of pieces of ribbon. To divide, I multiplied the whole number by the reciprocal of the fraction. $16 \div \frac{2}{5} = 16 \times \frac{5}{2} = \frac{80}{2} = 40$, so she made 40 pieces that each used $\frac{2}{5}$ yard of ribbon.</td>
</tr>
<tr>
<td><strong>Answer:</strong> Tom used more yards of ribbon.</td>
<td><strong>Work/Explanation:</strong> To find the length of ribbon Tom used, I multiplied the number of pieces Tom made by the length of each piece. $42 \text{ pieces } \times \frac{3}{7} \text{ yard} = \frac{126}{7} = 18$, so Tom used 18 yards of ribbon. Then I compared the lengths of ribbon Shirley and Tom used. Tom used 18 yards which is more than the 16 yards Shirley used. Or Other correct explanation</td>
</tr>
</tbody>
</table>

**2.0 score points**
1.0 point for correct answer and 1.0 point for all complete work and explanation or 0.5 point for enough correct work and/or explanation to follow what is being done (only minor omissions or errors)

**2.0 score points**
0.5 point for correct answer (or correct comparison based on incorrect answer carried through from Part A) and 1.5 points for all correct work and explanation or 1.0 point for enough correct work and/or explanation to follow what is being done (only minor omissions or errors) or 0.5 point for not enough correct work and/or explanation to follow what is being done or some correct work or explanation only

Figure 19: 4Sight Rubric Sixth Grade Question 29
**Problem 30** Figure A.2 depicts the second Grade 6 constructed response question. It asks students to demonstrate understanding of patterns, relations, and functions. It was described in detail in the body of the paper.
30. (Test 2) Nadia is learning about functions in her math class. She was given the assignment below for homework. Complete the assignment.

A. The function table below shows numbers that are related by a rule.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>96</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
</tbody>
</table>

Write the missing number in the empty space in the table.

B. Describe the rule for the function table in part A.

C. Make a new rule with a different operation. Use the rule to create a new function table with three pairs of numbers.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D. Describe the rule for the function table you created in part C.

30. (Test 2) Top Scoring Response

Part A

Answer:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>96</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
</tbody>
</table>

(1 score point)
1.0 point for correct answer

Part B

Explanation:
The rule is multiply by 12 or 12 times the x-value equals the y-value or \(12x = y\)

(1 score point)
1.0 point for correct explanation of the rule

Part C

Answer:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>60</td>
<td>6</td>
</tr>
<tr>
<td>90</td>
<td>9</td>
</tr>
</tbody>
</table>

(1.5 score points)
1.5 points for all correct values in the function table (that describes an addition, division or subtraction function)
or
1.0 point for enough correct work to follow what is being done (only minor omissions or errors)
0.5 point for not enough correct work to follow what is being done or some correct work only
Note: if the student creates a multiplication function table they cannot receive any points for part C

Part D

Explanation:
The rule is divide by 10 or the x-value divided by 10 gives the y-value or \(\frac{x}{10} = y\).
OR
Other correct explanation

(0.5 score point)
0.5 point for correct explanation of the rule in part C, even if student created a multiplication function

Figure 20: 4Sight Rubric Sixth Grade Question 30
A.3 Grade 8 rubrics

Problems 26 and 27 Both of the eighth grade constructed response problems use the same rubric form. The minor differences between them shown in Figure A.3 and Figure A.3 are simply the substitution of the correct answers for Problem 26 and 27 respectively. The points from each problem are split evenly between Part A and Part B. Each part splits its two points into a one point Yes/No judgment of the correct answer and a three level scale of 1 point, 0.5 points, or 0 points for the quality of the explanation.

26. (Test 4) Top Scoring Response

<table>
<thead>
<tr>
<th>Part A: Answer/Work</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% – 10% = 90%</td>
<td>The discounted price is less than the original price. Since the discount price is $12.50, you can work backwards to find the higher price. 100% – 10% = 90%, so I said $12.50 is 90% of the original price. I changed the 90% to 0.9 and divided $12.50 by 0.9 to find the original price.</td>
</tr>
<tr>
<td>$12.50 ÷ 0.9 = $13.8888</td>
<td>OR</td>
</tr>
<tr>
<td>or $13.89</td>
<td>Other correct explanation</td>
</tr>
<tr>
<td>$13.89 is the original price</td>
<td></td>
</tr>
</tbody>
</table>

(2 score points)

1.0 point for correct answer
and
1.0 point for all correct work and explanation
or
0.5 point for enough correct work and/or explanation to follow what is being done (minor omissions or errors present)

<table>
<thead>
<tr>
<th>Part B: Answer/Work</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 shirts = $33.00</td>
<td>The price of 1 discounted shirt is $12.50. The cost of the next 5 shirts is $11 × 4 (1 shirt is free with the special offer). The total price for the 6 shirts is $56.50. The 3% tax is calculated at $1.70 and added to the cost of the shirts to get the total amount of $58.20.</td>
</tr>
<tr>
<td>1 shirt free</td>
<td>$44.00</td>
</tr>
<tr>
<td>+ 12.50</td>
<td>$56.50</td>
</tr>
<tr>
<td>× 0.03</td>
<td>$1.6950 tax</td>
</tr>
<tr>
<td>+ 1.70</td>
<td>$58.20 total</td>
</tr>
<tr>
<td>1 shirt = $11.00</td>
<td>$56.50</td>
</tr>
<tr>
<td>5 shirts = $44.00</td>
<td>$58.20 total</td>
</tr>
</tbody>
</table>

(2 score points)

1.0 point for correct answer
and
1.0 point for all correct work and explanation
or
0.5 point for enough correct work and/or explanation to follow what is being done (minor omissions or errors present)

Figure 21: 4Sight Rubric Eighth Grade Problem 26
27. (Test 4) Top Scoring Response

<table>
<thead>
<tr>
<th>Part A: Answer/Work</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| \[
\frac{31}{52} = \frac{x}{312}, \quad 52x = 9672, \quad x = 186
\] | I can set up a proportion and solve for \(x\) to figure out about how many visitors Jason would expect to think the fruit dessert tasted best. Jason would expect 186 visitors to think the fruit dessert tasted best. |

(2 score points)
1.0 point for correct answer
and
1.0 point for all correct work and explanation
or
0.5 point for enough correct work and/or explanation to follow what is being done (minor omissions or errors are present)

<table>
<thead>
<tr>
<th>Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem-and-leaf plot AND Jason's data will show how frequently each age shows up in his survey. A stem-and-leaf plot is used to organize data into a frequency distribution.</td>
</tr>
</tbody>
</table>

(2 score points)
1.0 point for correct answer
and
1.0 point for a reasonable and correct explanation
or
0.5 point for enough reasonable explanation to follow the thinking (only minor omissions)