Do Learners Know What's Good for Them? Crowdsourcing Subjective Ratings of OERs to Predict Learning Gains

Jacob Whitehill
Worcester Polytechnic Institute, USA
jrwhitehill@wpi.edu

Cecilia Aguerrebere
Fundación Ceibal, Uruguay
caguerrebere@ceibal.edu.uy

Benjamin Hylak
Worcester Polytechnic Institute, USA
bhylak@wpi.edu

ABSTRACT

We explored[1] how learners' *subjective ratings* of open educational resources (OERs) in terms of how much they find them "helpful" can predict the actual *learning gains* associated with those resources as measured with pre- and post-tests. To this end, we developed a probabilistic model called GRAM (Gaussian Rating Aggregation Model) that combines subjective ratings from multiple learners into an aggregate quality score of each resource. Based on an experiment we conducted on Mechanical Turk ($n = 304$ participants with $m = 17$ math tutorial videos as resources), we found that aggregated subjective ratings are highly (and stat. sig.) predictive of the resources' average learning gains, with Pearson correlation of 0.78. Moreover, when predicting average learning gains of *new* learners, subjective scores were still predictive (Pearson correlation of 0.49) and attained higher prediction accuracy than a model that directly uses pre- and post-test data to estimate learning gains for each resource. These results have potential implications for large-scale learning platforms (e.g., MOOCs, Khan Academy) that assign resources (tutorials, explanations, hints, etc.) to learners based on the expected learning gains.

Keywords

open educational resources (OER); adaptive learning; crowdsourcing; treatment effect estimation

1. INTRODUCTION

Consider a hypothetical large-scale online learning platform in which learners engage with open educational resources (OERs) that are sampled from a vast collection. These resources could include tutorial videos, practice exercises, explanations of wrong answers, hints, etc. In order to help students learn optimally, the learning platform must decide which resource is most beneficial to each learner at each moment in time, and then assign that resource to the learner (see Figure 1). Although various criteria could be used for this decision (e.g., the impact on student engagement), perhaps the most natural one is how much the student will learn – *learning gains* – from receiving the resource.

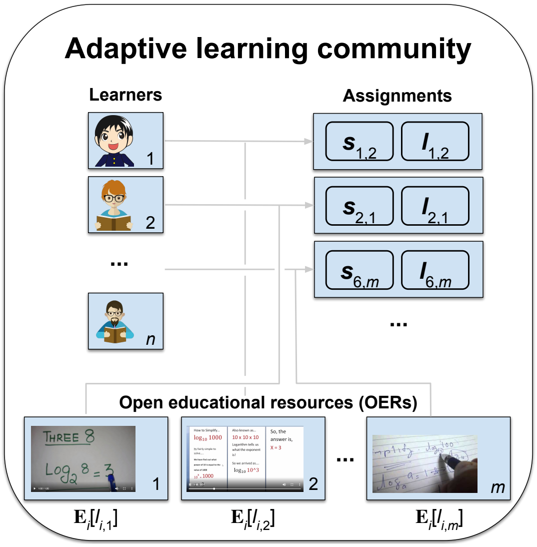


Figure 1: An adaptive learning community in which each learner i is assigned different resources over time, and the effectiveness (expected learning gains l_{ij}) of each resource j is estimated both from test scores as well as from *subjective ratings* s_{ij} given by the learners. Gray lines show hypothetical assignments of OERs to learners. $\mathbb{E}_i[l_{ij}]$ denotes the average learning gains over all learners i who received j.

The standard way to estimate the *learning gains* l_{ij} of each resource is to give each student i who receives resource j a pre-test (before receiving it) and post-test (after receiving it) to measure how much she/he learned, i.e., the difference between pre- and post-tests. We call each (learner, resource)-pair an *assignment.* After a sufficient number of assignments, the average learning gains of each resource $\mathbb{E}_i[l_{ij}]$ (averaged over all learners i who receive j) can be estimated. Then, using these estimates for all the resources, the most effective ones can be served to students. Unfortunately, this approach to estimating the quality of a large collection of

[1] The data and source code (in R) to reproduce the results in this paper are available at `https://github.com/jwhitehill/gram`.

OERs is expensive because testing takes a long time. On the other hand, after receiving a resource j, learners may have a *subjective opinion* about how effective j was. These opinions can arguably be queried more easily and efficiently than administering tests; for example, the learner could simply select between 1 and 5 stars (à la Yelp) to express how much she/he liked it. It is even possible that subjective scores might be better than test scores in some situations. For example, even if a learner her/himself has already mastered a skill and thus has a learning gain of 0, she/he might still be able to *judge* whether a resource is useful.

When using subjective scores to predict learning gains, care must be taken: some learners may be more or less reliable in making such judgments. However, there are reasons to be optimistic: (1) As long as enough learners "vote", then the noise of their judgments can be averaged out. (2) Using algorithms for crowdsourcing consensus (see below), the reliabilities of the learners as well as the learning gains of the resources can be estimated in an unsupervised fashion. The **chief contribution** of our work is to propose and evaluate experimentally an efficient crowdsourcing model to estimate the quality of a set of learning resources by combining multiple learners' subjective opinions about them.

2. RELATED WORK

Students' judgments of learning and teaching: Estimating the learning gains of an OER is related to metacognition. The ability of students to judge how well *other* people learn has been analyzed experimentally in prior works such as [12, 3]. However, we are not aware of previous research that considers this problem in the large scale of an online learning community or how to combine multiple learners' judgments to improve accuracy. In the context of student course evaluations, there is evidence that learners may actually be poor judges of their teachers' effectiveness [7, 4].

Adaptive online learning communities: Adaptive learning communities that decide which resources to serve to students based on up-to-date estimates have generated recent interest in the educational data mining and reinforcement learning communities. Notable works are by Rafferty, et al. [9] and Williams, et al. [17]. In these works, reinforcement learning techniques based on bandits and Thompson sampling were used both to estimate the learning gains of each resource and simultaneously to assign resources to learners. Our work is complementary: we explore how not only test score information, but subjective ratings provided by learners, could be useful in estimating the utility of each resource.

Crowdsourcing for education: In [14], Weld et al. provided an overview of how online learning creates challenges due to its large scale, but also suggests possible ways in which crowdsourcing can offer solutions to these challenges. Heffernan, et al. [6] proposes a vision of how crowdsourcing can help provide important functionality toward adaptive personalized online learning. As one specific instance of how the crowd can contribute new resources to an online learning community, Williams, et al. showed that people on Mechanical Turk can be induced to author novel and useful text-based explanations [17]. Whitehill & Seltzer [16] showed that Mechanical Turk workers can even create entire tutorial videos, at least some of which are effective at helping students to learn. Peer grading (e.g., Piech, et al. [8]) and peer feedback are other ways of harnessing the crowd to provide useful feedback for learners at scale.

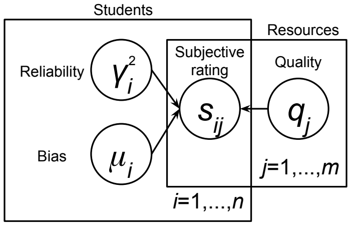


Figure 2: Gaussian Rating Aggregation Model (GRAM). Only the subjective ratings s_{ij} from student i about resource j are observed. Latent variable q_j expresses the "quality" of resource j and is used to predict the learning gains of students who receive the resource.

Crowdsourcing consensus algorithms: Since Dawid and Skene's seminal work [5] on optimal weighting of annotators' opinions, there have been a slew (e.g., [15, 10, 11, 2, 13, 1]) of crowdsourcing models, which are suitable for different kinds of tasks (binary, multiple choice, etc.) and capture different features of the labeling task (e.g., task difficulty, biases).

3. GAUSSIAN RATING AGGREGATION MODEL (GRAM)

We model the *quality* of each open educational resource (OER) j with a real number, q_j, that can be estimated by aggregating over many (real-valued) *subjective ratings* s_{ij} from many learners i. We thus develop a Gaussian probability model of how each s_{ij} is related to each q_j as well as several parameters specific to each learner i. The model is portrayed in Figure 2: Let μ_i and γ_i^2 be the bias and reliability (variance) of learner i, respectively. Let q_j be the ground-truth quality of resource j. We posit that student i's label s_{ij} for resource j is a Gaussian random variable with mean $q_j + \mu_i$ and variance γ_i^2. In other words, if the ground-truth quality is q_j, then student i adds a bias μ_i, and then adds independent 0-mean Gaussian noise with variance γ_i^2. We can express these relationships using the conditional probability density function (PDF) $P(s_{ij} \mid q_j, \mu_i, \gamma_i^2) = \mathcal{N}(q_j + \mu_i, \gamma_i^2)$ where $\mathcal{N}$ is a Gaussian with a given mean and variance.

3.1 Inference

As with many crowdsourcing consensus models, inference in the GRAM requires solving a "chicken-and-the-egg" problem: if the parameters μ_i, γ_i^2 of each learner i were known, then an optimal weighting of their votes s_{ij} could be used to estimate the quality q_j of each resource j. On the other hand, if the ground-truth quality q_j of each resource were known, then the parameters of each learner could be estimated. We solve this problem using Expectation-Maximization: in the E-Step we compute the PDF of each q_j conditional on the parameters $\{\mu_i, \gamma_i\}$. In the M-Step, we compute the expected joint log-likelihood of the $\{q_j\}$ and $\{s_{ij}\}$ w.r.t. the PDFs computed during the previous E-Step, and then maximize this expectation w.r.t. the parameters $\{\mu_i, \gamma_i\}$. Since the GRAM is Gaussian, both the E- and M-Steps can be done analytically, and thus the algorithm is very efficient.

Let the mean and variance of the prior distribution over each q_j be $\overline{m}$ and $\overline{s}^2$, respectively. Recall that the product of two Gaussian PDFs, with means m_1 and m_2 and variances s_1^2 and s_2^2, respectively, is also Gaussian and has a mean $s^2(m_1/s_1^2 + m_2/s_2^2)$ and variance $s^2 = (1/s_1^2 + 1/s_2^2)^{-1}$.

E-Step:

$$\begin{aligned}
\tilde{P}(q_j) &\doteq P(q_j \mid \{s_{ij}\}, \{\mu_i\}, \{\gamma_i^2\}) \\
&\propto P(q_j \mid \{\mu_i\}, \{\gamma_i^2\}) P(\{s_{ij}\} \mid q_j, \{\mu_i\}, \{\gamma_i^2\}) \\
&= P(q_j) \prod_i P(s_{ij} \mid q_j, \mu_i, \gamma_i^2) \\
&= \mathcal{N}(m_j, s_j^2) \quad \text{where} \\
s_j^2 &= \left(1/\overline{s}^2 + \sum_i 1/\gamma_i^2\right)^{-1} \\
m_j &= s_j^2 \left(\overline{m}/\overline{s}^2 + \sum_i (s_{ij} - \mu_i)/\gamma_i^2\right)
\end{aligned}$$

In other words, the posterior distribution of each q_j is a Gaussian whose mean is the average of the relevant s_{ij} after shifting each one by the learner's bias μ_i and then scaling it by γ_i^2. We can achieve a non-informative prior by setting the variance $\overline{s}^2$ to be very high (e.g., 1000).

M-Step: We derive the auxiliary function Q as the expectation, w.r.t. the PDF $\tilde{P}$ computed during the E-Step, of the joint log-likelihood of the observed ratings $\{s_{ij}\}$ and hidden ratings $\{q_j\}$. In the derivation below, C and D are constants that do not depend on any of the parameters.

$$\begin{aligned}
&Q(\{\mu_i\}, \{\gamma_i^2\}) \\
&\quad = \mathbb{E}\left[\log P(\{s_{ij}\}, \{q_j\} \mid \{\mu_i\}, \{\gamma_i^2\})\right] \\
&\quad = \mathbb{E}\Bigg[\log \prod_j P(q_j \mid \{\mu_i\}, \{\gamma_i^2\}) + \\
&\qquad\qquad \log \prod_{ij} P(s_{ij} \mid \{q_j\}, \{\mu_i\}, \{\gamma_i^2\})\Bigg] \\
&\quad = \mathbb{E}\left[\log \prod_j P(q_j) + \log \prod_{ij} P(s_{ij} \mid q_j, \mu_i, \gamma_i^2)\right] \\
&\quad = \sum_j \mathbb{E}[\log P(q_j)] + \sum_{ij} \mathbb{E}[\log P(s_{ij} \mid q_j, \mu_i, \gamma_i^2)] \\
&\quad = \sum_{ij} \int_{-\infty}^{+\infty} dq_j \tilde{P}(q_j)[\log P(s_{ij} \mid q_j, \mu_i, \gamma_i^2)] \ + C \\
&\quad = -\sum_{ij} \int_{-\infty}^{+\infty} dq_j \tilde{P}(q_j) \left[\frac{(s_{ij} - q_j - \mu_i)^2}{2\gamma_i^2} + \log \gamma_i\right] \ + D \\
&\quad = -\sum_i \log \gamma_i \ - \\
&\qquad\quad \frac{1}{2} \sum_{ij} \int_{-\infty}^{+\infty} dq_j \tilde{P}(q_j)[(s_{ij} - q_j - \mu_i)^2/\gamma_i^2] \ + D \\
&\quad = -\sum_i \log \gamma_i - \frac{1}{2} \sum_{ij} \Bigg[(s_{ij} - \mu_i)^2/\gamma_i^2 \ - \\
&\qquad\quad \frac{2(s_{ij} - \mu_i)}{\gamma_i^2} \int_{-\infty}^{+\infty} dq_j \tilde{P}(q_j) q_j + \frac{1}{\gamma_i^2} \int_{-\infty}^{+\infty} dq_j \tilde{P}(q_j) q_j^2\Bigg]
\end{aligned}$$

where we omitted the constant D in the last line for brevity. The two integrals are the first and second plain moments of $\tilde{P}(q_j)$. The first is the mean of $\tilde{P}(q_j)$, i.e., m_j. The second can be obtained using the fact that the variance $\mathbb{V}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ for any random variable x. The second plain moment is thus $m_j^2 + s_j^2$. Hence,

$$\begin{aligned}
&Q(\{\mu_i\}, \{\gamma_i^2\}) = -\sum_i \log \gamma_i \ - \\
&\qquad \frac{1}{2} \sum_{ij} \frac{1}{\gamma_i^2} \left[(s_{ij} - \mu_i)^2 - 2(s_{ij} - \mu_i) m_j + m_j^2 + s_j^2\right]
\end{aligned}$$

We now differentiate with respect to each parameter, set to 0, and solve:

$$\begin{aligned}
\frac{\partial Q}{\partial \mu_i} &= -\frac{1}{2} \frac{1}{\gamma_i^2} \sum_j (-2(s_{ij} - \mu_i) + 2m_j) \\
0 &= -\frac{1}{\gamma_i^2} \sum_j (\mu_i - s_{ij} + m_j) \\
\sum_j \mu_i &= \sum_j (s_{ij} - m_j) \\
\mu_i &= \frac{1}{N_i} \sum_j (s_{ij} - m_j) \quad \text{where} \qquad (1) \\
& \quad\; N_i \text{ is the \# of ratings from person } i \\
\frac{\partial Q}{\partial \gamma_i} &= -1/\gamma_i \ + \\
& \quad\; \frac{1}{\gamma_i^3} \sum_j \left[(s_{ij} - \mu_i)^2 - 2(s_{ij} - \mu_i) m_j + m_j^2 + s_j^2\right] \\
\gamma_i^2 &= \sum_j \left[(s_{ij} - \mu_i)^2 - 2(s_{ij} - \mu_i) m_j + m_j^2 + s_j^2\right] \ (2)
\end{aligned}$$

For our experiments we conducted 50 EM iterations.

3.2 Regularizing the model

In the full-fledged GRAM, all of the parameters (bias and reliability of each rater) are learned in an unsupervised fashion (see Section 3.1). Given enough data, these parameters can lead to more accurate estimates of each q_j. However, given limited data, it can also be useful to regularize the model by removing parameters and/or fixing them to known values. In fact, if there are too few subjective scores s_{ij} per learner, then it is important to remove some parameters because otherwise the model encounters identifiability problems. Hence, we considered several variants of the GRAM: (1) each γ_i^2 is estimated, but each μ_i is fixed to 0; (2) each μ_i is estimated, but $\gamma_i^2 = 1$. Finally, we also explored the hypothesis that the students with the higher pre-test scores might, perhaps due to a higher overall engagement, also be more reliable in giving subjective ratings. Hence, we also tried: (4) μ_i is estimated, but $\gamma_i^2 = 1/\sqrt{\mathbb{E}_j[p_{ij}] + \epsilon}$, where $\mathbb{E}_j[p_{ij}]$ is the average (over all their assignments) pre-test score p_{ij} of student i before receiving resource j, and $\epsilon = 0.1$ ensures that the denominator is positive.

4. MODELS FOR COMPARISON

We compared the GRAM to two other models: (1) unweighted average of subjective scores, and (2) prediction model trained directly on pre- and post-test scores.

4.1 Unweighted average of subjective scores

Instead of using the GRAM, we can estimate the quality q_j of each resource j simply as the unweighted average, over all learners who rated j, of their subjective rating scores s_{ij}.

4.2 Average post-test minus pre-test scores

The primary goal of our paper is to assess to what extent subjective scores can estimate the learning gains as measured in a pre-test/post-test paradigm. Hence, a strong baseline – indeed, a likely upper bound – to which to compare our GRAM approach is using a prediction model that *directly* uses test scores (on training data) to estimate students' learning gains (on testing data). In particular, for each resource j, we estimate $\mathbb{E}_i[l_{ij}]$ – the average difference between post-tests and pre-tests of all students i in the training set who received resource j. We then use this number to predict the average learning gains of resource j in the test set. Obviously, this requires that the adaptive learning system administer pre- and post-tests to learners in order to assess each resource's quality, and this can be much more time-consuming than simply asking the learner how much she/he likes it. Note that we also considered a prediction model that additionally uses students' pre-test scores as a co-variate, which could model possible ceiling effects in the tests. However, our results with that model were slightly worse, and hence we do not report them.

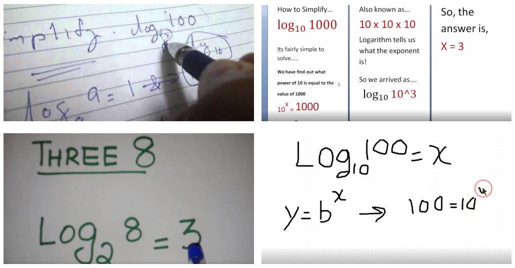
5. EXPERIMENT

To assess how well subjective scores of the resources' quality predicted their associated learning gains, we conducted a randomized expeirment on Mechanical Turk. Each participant was paid $1 and could complete up to 3 tasks. In each task, the pre- and post-tests were the same, but the learning resource was usually different due to random assignment.

5.1 Overview

During the task, participants learned about logarithms. Logarithms are a topic that many adults have learned, but many have forgotten. The topic is hard enough to induce variability in test scores, but easy enough to be learned (or refreshed) in a short amount of time. The learning resources in our experiment comprised a set of tutorial videos on logarithms, most of which were 2-3 minutes long. These resources were authored by different people around the world and collected in a study by Whitehill & Seltzer [16]. Each tutorial explains the solution to one of the math problems that appeared on the pre-test (see Figure 4).

To select videos for our experiment, we watched over 100 candidate tutorial videos collected by [16]. Each video was watched by at least one of the investigators and labeled as either "High Quality," "Low Quality," or "Not Acceptable." Videos labeled as "Not Acceptable" were excluded. To induce some variability in the quality of videos, we chose one "High Quality" video as well as one "Low Quality" for each of the Basic Logarithm problems in the pre-test (see Figure 4), except for a few problems where only one quality level was available. In total, there were $m = 17$ resources (tutorials) that could be assigned; see Figure 3 for examples.



How to Simplify...

$\log_{10} 1000$

Its fairly simple to solve.....

We have find out what power of 10 is equal to the value of 1000

$10^x = 1000$

Also known as...

10 x 10 x 10

Logarithm tells us what the exponent is!

So we arrived as....

$\log_{10}$ 10^3

So, the answer is,

X = 3

Figure 3: Sample learning resources (tutorial videos on logarithms from [16]) that we used in our experiment.

Basic Logarithms – Simplify:	
$\log_3 1 =$	$\log_9 1 =$
$\log 100 =$	$\log_{\frac{1}{5}} 125 =$
$\log_{10} 1000 =$	$\log_{\frac{1}{x}} x^2 =$
$\log_3 81 =$	$\log_w \frac{1}{w} =$
$\log_2 8 =$	$\log_{\frac{1}{2}} \frac{1}{4} =$
Logarithms and Variables – Simplify:	
$\log_a a^2 =$	$\log_x x^4 =$
$\log_4 4^{2b} =$	$\log_{x-1}(x-1)^y =$
Equations with Logarithms – Solve:	
$\log_3(x-1) = 4$	$x \log_4 16 = 3$
$z \log_{10} \sqrt{10} = 4$	$y \log_{10} 1000 = 3$

Figure 4: The pre-test on logarithms (borrowed from [16]) in our experiment.

5.2 Protocol

The experiment was built as a web application using HTML and Javascript. Each session consisted of multiple phases:

1. **Survey**: The participants were first asked some basic demographic questions, such as their highest level of education, gender, and age. (Note that we did not use these data in the analyses in this paper.)

2. **Pre-test**: The pre-test surveyed their pre-existing skills in three areas: Basic Logarithms, Logarithms and Variables, and Equations with Logarithms.

3. **Tutorial video**: Participants were then randomly assigned one of the 17 different tutorial videos.

4. **Subjective rating of the resource**: On a Likert scale of 1 to 5, participants were asked how much they agreed with the statement: "This video will help other students learn about logarithms."

5. **Post-test**: The post-test contained different math problems but was otherwise comparable in format, subject matter, and difficulty to the pre-test.

6. RESULTS AND ANALYSIS

A total of $n = 304$ participants completed the task. Of these, 239 completed 1 task, 35 completed 2 tasks, and 30 completed 3 tasks. Figure 5 shows the box plot, for each resource (tutorial video) j, of the learning gains associated with each resource. There is high variance in learning gains *within* each resource ($\mathbb{E}_j[\mathbb{V}_i[l_{ij}]]$ averaged over the $m = 17$

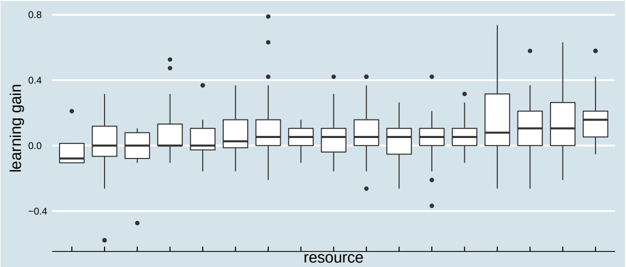


Figure 5: Box plot, for each OER (math tutorial video) j, of the learning gains l_{ij} for all users i who received it. Resources are sorted according to their median learning gains over all learners who received them.

videos is 0.03) that dwarfs the variance in average learning gains *between* resources ($\mathbb{V}_j[\mathbb{E}_i[l_{ij}]]$ is 0.003), where $\mathbb{V}_i[\cdot]$ denotes the variance with respect to learners i and $\mathbb{V}_j[\cdot]$ denotes the variance with respect to resources j. The relative magnitudes of these variances makes the prediction of average learning gains $\mathbb{E}_i[l_{ij}]$ of each individual resource a challenging task.

6.1 Are subjective ratings correlated with average learning gains?

From the set of all subjective scores collected in our experiment, we can aggregate the ratings s_{ij}, using either the proposed GRAM or simply the unweighted average, for each resource j into a quality estimate q_j. Similarly, we can compute the average learning gains associated with each resource j over all students assigned j to obtain $\mathbb{E}_i[l_{ij}]$, where the subscript indicates that the expectation is w.r.t. all students assigned j. This is equivalent to estimating the *average treatment effect* of resource j. We then compute the correlation (Pearson, Spearman) between these two sets of variables. Note that, since many learners in our experiment completed only one task, we needed to simplify the GRAM in order to avoid identifiability problems (see Section 3.2). Hence, instead of the full-fledged GRAM, we used two variants: one where each $\mu_i = 0$, and one where $\mu_i = 0$ and γ_i^2 is determined by the learner's pre-test score.

Results are shown in Table 1. Because all correlations are estimated within-sample (i.e., there is no separation of training and validation data), computing the p-values (two-tailed) is straightforward. When the GRAM was used to infer only the reliability γ_i^2 (first line of Table 1), the accuracy is low – 0.15 (Pearson) and 0.13. On the other hand, with the other two GRAM variants, when either a bias μ_i for each labeler is learned, the performance was much better – up to 0.78 (Pearson) and 0.75 (Spearman) between the inferred q_j and the average learning gains. These results are easily better than what is obtained using just the unweighted average of the learners' ratings. Estimating γ_i^2 as a function of each learner's pre-test score did not yield a clear accuracy improvement. Altogether, the results suggest that, with the right aggregation model, learners' subjective scores carry considerable information about the average learning gains of the resources they receive.

Predicting learning gains within-sample

Method	Pearson	Spearman
GRAM (learn γ_i^2)	0.15 ($p = 0.56$)	0.13 ($p = 0.63$)
GRAM (learn μ_i)	0.78 ($p < 0.001$)	0.70 ($p = 0.002$)
GRAM (learn μ_i, set γ_i^2 from pretest)	0.76 ($p < 0.001$)	0.75 ($p < 0.001$)
Unweighted average	0.38 ($p = 0.14$)	0.54 ($p = 0.03$)

Table 1: Accuracies, and associated p-values, of different models when predicting the average learning gains $\mathbb{E}_i[l_{ij}]$ of the resources from subjective ratings reported by learners. For aggregating learners' subjective ratings, we consider both the unweighted average as well as the quality scores inferred using the GRAM.

6.2 Do subjective ratings predict the average learning gains for new students?

Suppose some *new* students enter the adaptive learning community. How accurately can we predict the average learning gains $\mathbb{E}_i[l_{ij}]$ of a resource j for these learners? How does this accuracy compare to that of a prediction model in which we estimate the effectiveness of each resource directly based on pre- and post-test data?

We conducted 3-fold cross-validation, where the same students never appear in more than one fold. From the training data in each fold, we use GRAM to infer the latent variables q_j from the subjective scores s_{ij}; we use the variant in which only μ_i is learned. We then compute the correlation (Pearson, Spearman) between q_j and the average learning gains of resource j over all students i in the test set who received j. Due to the high variability in results over the 3 folds, we repeated the 3-fold cross-validation 30 times, and averaged the results over trials. In each trial, we ensured that the data were randomly partitioned such that every resource was assigned to at least 1 learner in at least 2 folds (i.e., one testing fold and one training fold).

In the cross-validation framework, computing p-values is not straightforward because the estimates from each fold are not statistically independent. Instead, we estimated the uncertainty of each correlation as the average (over the 30 trials) standard error (i.e., the standard deviation of the correlations over the $K = 3$ folds, divided by $\sqrt{K}$). We compare the accuracy of predictions obtained with the GRAM to the predictions by the *unweighted average* model (Section 4.1), and also to the predictions from a model that has direct access to the *test scores* (see Section 4.2). The latter is a strong comparison because it has access to actual pre- and post-test scores, whereas the other models do not.

Results are shown in Table 2. The GRAM – which utilizes only subjective scores, not test results, of the training data – is able to predict the average learning gains for new learners with higher accuracy (0.49 Pearson and 0.43 Spearman correlation) compared to the model that uses pre- and post-test data (0.36 Pearson and 0.41 Spearman correlation) to estimate the quality of each resource. Even the unweighted average of learners' subjective ratings retains most of the prediction accuracy that could be achieved using explicit pre-

Predicting learning gains for new students

Method	**Pearson**	**Spearman**
GRAM (learn μ_i)	0.49 (±0.11)	0.43 (±0.11)
Unweighted average	0.32 (±0.09)	0.35 (±0.11)
Predict from test scores	0.36 (±0.09)	0.41 (±0.08)

Table 2: Accuracies (± their standard errors) over $K = 3$ cross-validation folds, of different models when predicting the average learning gains $\mathbb{E}_i[l_{ij}]$ of *new* learners (i.e., not used for training).

and post-test score data. All in all, our results suggest that (1) learners' subjective ratings carry considerable information that could be useful in an adaptive learning community for deciding which resources are more effective than others, and (2) using a crowdsourcing consensus model such as our proposed GRAM can potentially yield higher accuracy than simply taking the unweighted average.

7. CONCLUSION

We investigated whether learners' subjective opinions about the quality of learning resources (e.g., a tutorial video) are correlated with the learning gains (post-test minus pre-test) associated with receiving those resources. This could have implications for adaptive online learning communities in which open educational resources (OER) are served to students based on estimates of how effective they would be for learning: Rather than giving relatively time-consuming pre- and post-tests, the adaptive learning platform could instead simply ask learners how helpful they found the resources to be. We developed a novel Gaussian Rating Aggregation Model (GRAM) with which to aggregate many learners' subjective scores into an overall quality estimate for each resource. Based on an experiment that we conducted on Mechanical Turk, we found that (1) subjective scores are highly correlated with average learning gains (Pearson correlation of 0.78). Moreover, (2) when predicting the average learning gains for learners who are *new* to the learning community, the accuracy (Pearson correlation of 0.49) using the GRAM from subjective scores was even better than estimating learning gains from test scores.

Future work will consider how to combine subjective scores with test data in order to arrive at improved estimation accuracy of resources' effectiveness. Moreover, with the goal to personalize education, it would be interesting to explore how to harness subjective ratings to estimate *individual* learning gains rather than just *average* learning gains. Finally, it is important to establish whether the results we collected in our study on adult participants from Mechanical Turk generalizes to more authentic online learning communities (e.g., Khan Academy, ASSISTments).

Acknowledgement and disclaimer: We gratefully acknowledge a grant (N00014-18-1-2768) from the Office of Naval Research that supported this research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

8. REFERENCES

[1] Y. Baba and H. Kashima. Statistical quality estimation for general crowdsourcing tasks. In *Knowledge discovery and data mining*. ACM, 2013.

[2] J. Bragg, D. S. Weld, et al. Crowdsourcing multi-label classification for taxonomy creation. In *AAAI conf. on human computation and crowdsourcing*, 2013.

[3] R. Bromme, R. Rambow, and M. Nückles. Expertise and estimating what other people know: The influence of professional experience and type of knowledge. *Journal of experimental psychology: Applied*, 2001.

[4] S. E. Carrell and J. E. West. Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 2010.

[5] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, 1979.

[6] N. T. Heffernan, K. S. Ostrow, K. Kelly, D. Selent, E. G. Van Inwegen, X. Xiong, and J. J. Williams. The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education*, 26(2), 2016.

[7] P. A. Kirschner and J. J. van Merriënboer. Do learners really know best? urban legends in education. *Educational psychologist*, 48(3), 2013.

[8] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*, 2013.

[9] A. N. Rafferty, H. Ying, and J. J. Williams. Bandit assignment for educational experiments: Benefits to students versus statistical power. In *Artificial Intelligence in Education*, 2018.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 2010.

[11] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, 1995.

[12] J. G. Tullis. Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & cognition*, 46(8):1360–1375, 2018.

[13] R. K. Vinayak and B. Hassibi. Crowdsourced clustering: Querying edges vs triangles. In *Advances in Neural Information Processing Systems*, 2016.

[14] D. S. Weld, E. Adar, L. Chilton, R. Hoffmann, E. Horvitz, M. Koch, J. Landay, C. H. Lin, and M. Mausam. Personalized online education–a crowdsourcing challenge. In *Workshops at AAAI*, 2012.

[15] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, 2009.

[16] J. Whitehill and M. Seltzer. A crowdsourcing approach to collecting tutorial videos–toward personalized learning-at-scale. In *Learning@ Scale*. ACM, 2017.

[17] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Learning@ Scale*. ACM, 2016.