A Crowdsourcing Approach to Collecting Tutorial Videos – Toward Personalized Learning-at-Scale

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

Margo Seltzer
Harvard University
Cambridge, MA, USA
margo@eecs.harvard.edu

ABSTRACT

We investigated the feasibility of crowdsourcing full-fledged tutorial videos from ordinary people on the Web on how to solve math problems related to logarithms. This kind of approach (a form of *learnersourcing* [9, 11]) to efficiently collecting tutorial videos and other learning resources could be useful for realizing *personalized learning-at-scale*, whereby students receive specific learning resources – drawn from a large and diverse set – that are tailored to their individual and time-varying needs. Results of our study, in which we collected 399 videos from 66 unique "teachers" on Mechanical Turk, suggest that (1) approximately 100 videos – over 80% of which are mathematically fully correct – can be crowdsourced per week for $5/video; (2) the average learning gains (posttest minus pretest score) associated with watching the videos was stat. sig. higher than for a control video (0.105 versus 0.045); and (3) the average learning gains (0.1416) from watching the best tested crowdsourced videos was comparable to the learning gains (0.1506) from watching a popular Khan Academy video on logarithms.

INTRODUCTION & RELATED WORK

The goal of *personalized learning*, in which students' learning experiences are tailored to their individual and time-varying needs, has been been pursued by psychologists, computer scientists, and educational researchers for over five decades [15, 2, 17, 4, 8, 5]. A key challenge when developing personalized learning systems is *how to efficiently collect a set of learning resources* – e.g., illuminating tutorials of key concepts, edifying practice problems, helpful explanations of how to solve them, etc. – that are used to personalize instruction [1]. Without a sufficiently large and diverse set of resources from which to draw, personalized learning may not offer much advantage over traditional, single-path instruction.

L@S 2017, April 20 - 21, 2017, Cambridge, MA, USA

ISBN 978-1-4503-4450-0/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3051457.3053973

One recently proposed and promising approach to collecting and curating large volumes of educational resources is to *crowdsource* data from learners themselves. This process, sometimes known as *learnersourcing*, has been used, for example, to identify which parts of lecture videos are confusing [9], and to describe the key instructional steps [11] and subgoals [10] of "how-to" videos. More recently, learnersourcing has been used not only to annotate existing educational content, but also to create novel content itself [6, 12, 16].

In this paper, we too explore a crowdsourcing approach to efficiently collecting a large and diverse set of learning resources. However, in contrast to previous work, which thus far has focused on text-based content, our work is concerned with asking ordinary people from a crowdsourcing web site to take on the role of a teacher ("teachersourcing" [7]) and to create *novel, full-fledged, video-based explanations* that provide worked examples [3] of how to solve math problems. In contrast to static text, multimedia videos such as whiteboard animations can help to focus students' attention on the most salient parts of an explanation – e.g., by pointing to a specific mathematical expression with the mouse pointer while talking. Moreover, some students may find video to be more engaging than text, and there is preliminary evidence from the education literature that multimedia presentations lead to greater knowledge retention compared to static text-based presentations [14]. We note that the effort involved for the "teachers" in creating these videos is considerable – often an hour or more according to self-reports by the participants in our study. It is thus unclear how many people on crowdsourcing websites such as Mechanical Turk would even respond to such a task, and even less clear how useful such crowdsourced explanations might be for helping students to learn.

This paper describes what we believe to be the first investigation into crowdsourcing entire tutorial videos from ordinary people on the Web. In particular, the rest of the paper investigates two main research questions: (1) How feasible is it to attempt to crowdsource novel (i.e., not just a link to an existing video), full-fledged tutorial videos of math concepts from ordinary people on a crowdsourcing website (e.g., Mechanical Turk)? (2) How effective are these videos in helping students to learn? Further results and analyses are available at https://arxiv.org/pdf/1606.09610.pdf.

EXPERIMENT 1: CROWDSOURCING VIDEOS

We focused on crowdsourcing tutorial videos that explain how to simplify mathematical expressions and solve equations involving *logarithms*. Logarithms are well-suited for this study because (1) many people know what they are; (2) many other people – even those who once learned them many years ago – do not; and (3) people who are not familiar with logarithms can still learn something useful about them in a small (< 10 minutes) amount of time. In particular, we chose 18 math problems related to logarithms – e.g., "Simplify $\log_3 81$", or "Solve for x: $\log_3(x-1) = 4$" – that were given as part of a pre-test from another research project [13] on how math tutors interact with their students in traditional classroom settings.

Participants: The "teachers" in our study were adult (18 years or older) workers on Amazon Mechanical Turk. All participants were first required to give informed consent (Harvard IRB15-0867) and also sign a video recording release form so that their video explanations could be used in subsequent experiments on learning. Participants who completed the experiment received a payment of \$5. **Procedure**: In order to give the participant an idea of what we were looking for, we asked her/him to watch several examples of what a good video explanation might look like; the examples we chose were popular videos from YouTube about long division and quadratic equations. For the benefit of participants who chose to record their own handwriting, we also provided explicit guidelines on handwriting quality and showed good and bad examples of each. After presenting these guidelines, we showed a particular math problem – e.g., "Simplify $\log_3 81$" – and asked them to create and upload a video explaining how to solve it. **Dependent variables** were (1) the number of participants who created a tutorial video, (2) the average number of tutorial videos created by each participant, and (3) the fraction of submitted videos that were mathematically correct.

Results

Over 2 data collection periods consisting of approximately 2 weeks each, we collected 399 videos from 66 unique teachers (17% female; minimum reported age of 18, maximum reported age of 55) – approximately 6 videos per participant. This corresponds to approximately 100 videos per week of active data collection. The duration of most videos was between 1 and 3 minutes. Interestingly, several of the participants in our study expressed to us via email their enjoyment in completing the HIT, and many of them created explanations for several different problems. See Figure 1 for a small sample of the crowdsourced videos.

Analysis of correctness

To-date, we have manually annotated 145 out of the 399 submissions for mathematical correctness. Of the 145 annotated videos, 117 (81% of 145) were judged to be fully correct (i.e., contained no objectively false assertions); 16 videos (11%) contained at least one mathematical error; 7 (5%) were judged as "borderline" (contained minor missteps such as when the teacher referred to a mathematical *expression* (e.g., $\log_2 1$) as an *equation*); and 5 (3%) were not proper submissions (e.g., the submission was not a math video at all).

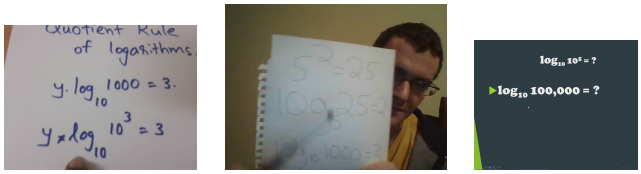

Figure 1. Snapshots of 3 representative examples of 399 total crowdsourced explanatory videos on logarithms.

EXPERIMENT 2: FINDING THE BEST VIDEOS

After crowdsourcing the videos, we next explored whether the videos show any promise for actually helping students to learn. Because this study is about crowdsourcing novel explanations from ordinary people around the world who may have varying mathematical skill and pedagogical expertise, we do not expect *all* the videos to be effective in helping students to learn. Rather, we assessed whether the *average* learning effectiveness of the videos – quantified by posttest-minus-pretest score of participants who watched the videos in a separate experiment – was statistically significantly higher than the learning effectiveness of a "control" video about a math topic unrelated to logarithms (specifically, a historical tutorial about the number π).

With this goal in mind, we randomly sampled 40 videos from the 117 that were confirmed (out of the 145 total that were annotated) to be mathematically correct. We then used these videos to conduct the experiment described below. In contrast to Experiment 1, the participants in this experiment were not expected to know anything *a priori* about logarithms.

Participants: We recruited $N = 200$ participants from Amazon Mechanical Turk. Each participant who completed the experiment received \$0.40 payment. **Apparatus**: We created a Web-based pretest and posttest on logarithms based on the materials in [13]. **Procedure**: After taking a pretest, each participant watched a randomly assigned video – with probability 0.2, the participant was assigned the control video, and with uniform probability of $0.8/40 = 0.02$, the participant was assigned to watch one of the 40 crowdsourced videos. Then the participant took a posttest. The **dependent variables** in this experiment were the average learning gains

$$G_k \doteq \frac{1}{|V(k)|} \sum_{i \in V(k)} (\text{post}_i - \text{pre}_i)$$

for each video k, where pre_i and post_i are the pretest and posttest scores (each between 0 and 1) for participant i, and $V(k)$ is the set of participants who were assigned to watch video k.

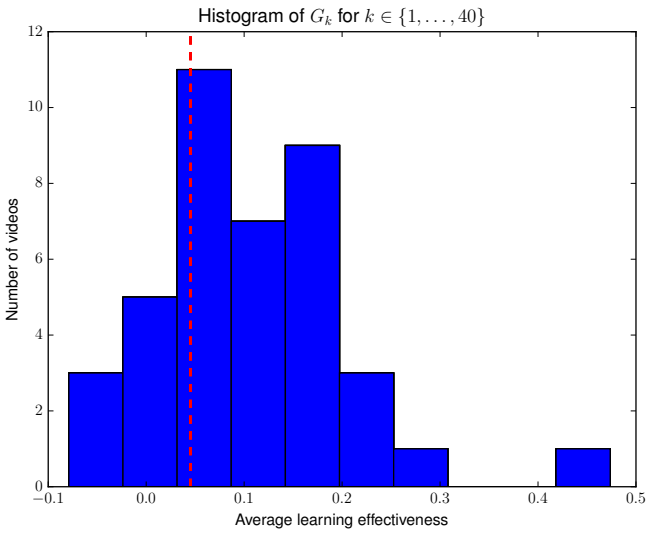

Figure 2. Histogram of the average learning gains G_k (average posttest minus pretest score across all subjects who watched video k) for 40 ($k \in \{1, \ldots, 40\}$) crowdsourced videos. The red dashed line shows the average learning gains for the "control" video.

Results

The histogram of the G_k for $k \in \{1, \ldots, 40\}$ is shown in Figure 2. The average learning gains (0.105) for the 40 crowdsourced videos that were tested was higher than for the control video (0.045); the difference was statistically significant ($t(39) = 3.715$, $p < 0.001$, two-tailed).

Differential Drop-out

Since some subjects started but did not complete the experiment, the number of subjects collected per video varied. This issue of differential drop-out can lead to distorted estimates: for example, if one tutorial video is particularly bad and only those students who are highly engaged decide to persist through the bad video and complete the HIT, then the estimated learning gains for that video might be positively biased. Unfortunately, Mechanical Turk does not provide an easy mechanism to track which workers started, but did not complete, the experiment – data are available only for participants who finished the entire HIT. However, since we do know how many participants completed the HIT for each video, and since we know the prior probability of assigning each participant to each video, we can assess whether some videos resulted in drop out more often than others. Specifically, we conducted a Pearson's χ^2 test where the vector of probabilities for the 41 videos (1 control plus 40 crowdsourced videos) was $[0.2\ \frac{0.8}{40}\ \ldots\ \frac{0.8}{40}]$. The result of the test ($\chi^2(40) = 34$, $p = 0.7363$) indicate that the *completion* rates for the videos were not statistically significantly different from the *assignment* rates. Though this result does not mean that the estimates of learning effectiveness in Figure 2 are completely unbiased, it provides some evidence that they are not to be completely discounted.

Video	Participants	G_k
1	58	0.1416
2	42	0.1140
3	57	0.0942
4	35	0.0932
Khan	58	0.1506

Table 1. Average learning gains G_k as measured in Experiment 3, for the 4 videos were estimated to be highest in Experiment 2, compared to the average learning gains of a popular Khan Academy video on logarithms.

EXPERIMENT 3: COMPARING TO KHAN ACADEMY

In our third experiment, we compared the learning gains of the best 4 videos as estimated in Experiment 2, to the learning gains of a popular tutorial video on logarithms produced by Khan Academy (https://www.youtube.com/embed/Z5myJ8dg_rM, with 924,520 views as of October 20, 2016). **Participants**: We recruited $N = 250$ participants from Amazon Mechanical Turk. Each participant who completed the experiment received $0.40 payment. **Apparatus**: Same as in Experiment 2. **Procedures**: Same as in Experiment 2 except that each participant was assigned uniformly at random to watch one of five different tutorial videos: 4 of these videos were crowdsourced videos, and 1 was the Khan Academy video. The **dependent variables** were the same as in Experiment 2.

Results

As shown in Table 1, the learning gains associated with the Khan Academy video compared to the best of the 4 crowdsourced videos were very similar – 0.1506 versus 0.1416, respectively. The difference between them was not statistically significant ($t(114) = 0.2277, p = 0.82$, two-tailed).

We note the following issues when comparing the crowdsourced math videos to the Khan Academy video: On the one hand, the Khan Academy video was substantially longer (7 minutes and 2 seconds) than the 4 crowdsourced videos (maximum length 2 minutes and 16 seconds) and hence can contain subtantially more potentially useful math content. On the other hand, the content presented in the crowdsourced videos was arguably more closely aligned to the post-test (though none of the questions explained in the video was exactly the same as any problem on the post-test) than was the Khan Academy video. Nonetheless, the results suggest that math tutorials crowdsourced from ordinary people on the Web can, at least sometimes, produce high-quality educational content.

SUMMARY & CONCLUSIONS

We explored how to devise a crowdsourcing task for use on Amazon Mechanical Turk in which ordinary people are asked to take on the role of a "teacher" and create novel tutorial videos that explain how to solve specific math problems related to logarithms. Results of three experiments suggest that crowdsourcing of full-fledged tutorial videos from ordinary people is feasible, provided

that appropriate guidelines (e.g., about using clear handwriting) on how to craft the explanations are provided. In fact, several of the crowdsourced workers expressed enthusiasm for the task, which likely requires more creativity than the kinds of tasks that are typically crowdsourced (e.g., image tagging). Although a few of the crowdsourced tutorial videos – which would need to be filtered out – contained important mathematical errors, the best of these videos were statistically significantly more effective, in terms of helping students to learn, than what would be expected from a "control" video on an irrelevant math topic. In fact, in terms of associated learning gains, the very best crowdsourced videos were comparable – and statistically indistinguishable from – a popular tutorial video on logarithms produced produced by Khan Academy. In sum, these findings provide support for the hypothesis that crowdsourcing can play an important role in collecting large, rich, and diverse sets of educational resources that enable personalized learning at scale.

Acknowledgement

The authors gratefully acknowledge a Spark grant (Spring 2015) from the Harvard Institute of Learning and Teaching (HILT) that partially funded this research.

REFERENCES

1. T. Aleahmad, V. Aleven, and R. Kraut. Creating a corpus of targeted learning resources with a web-based open authoring tool. *IEEE Transactions on Learning Technologies*, 2(1):3–9, 2009.
2. J. R. Anderson, C. F. Boyle, and B. J. Reiser. Intelligent tutoring systems. *Science*, 228(4698):456–462, 1985.
3. J. L. Booth, K. E. Lange, K. R. Koedinger, and K. J. Newton. Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction*, 25:24–34, 2013.
4. P. Brusilovsky and C. Peylo. Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education (IJAIED)*, 13:159–172, 2003.
5. C.-M. Chen. Intelligent web-based learning system with personalized learning path guidance. *Computers & Education*, 51(2):787–814, 2008.
6. Y. Chen, T. Mandel, Y.-E. Liu, and Z. Popovic. Crowdsourcing accurate and creative word problems and hints. 2016.
7. N. T. Heffernan, K. S. Ostrow, K. Kelly, D. Selent, E. G. 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*, pages 1–30, 2016.
8. G.-J. Hwang, F.-R. Kuo, P.-Y. Yin, and K.-H. Chuang. A heuristic algorithm for planning personalized learning paths for context-aware ubiquitous learning. *Computers & Education*, 54(2):404–415, 2010.
9. J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of Learning at Scale*, pages 31–40. ACM, 2014.
10. J. Kim, R. C. Miller, and K. Z. Gajos. Learnersourcing subgoal labeling to support learning from how-to videos. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 685–690. ACM, 2013.
11. J. Kim, P. T. Nguyen, S. Weir, P. J. Guo, R. C. Miller, and K. Z. Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 4017–4026. ACM, 2014.
12. O. Polozov, E. O'Rourke, A. M. Smith, L. Zettlemoyer, S. Gulwani, and Z. Popovic. Personalized mathematical word problem generation. In *IJCAI*, pages 381–388, 2015.
13. L. P. Salamanca, A. R. Carini, M. A. Lee, K. Dykstra, J. Whitehill, D. Angus, J. Wiles, J. S. Reilly, and M. S. Bartlett. Characterizing the temporal dynamics of student-teacher discourse. In *International conference on Development and Learning*, pages 1–2, 2012.
14. S. Türkay. The effects of whiteboard animations on retention and subjective experiences when learning advanced physics topics. *Computers & Education*, 98:102–114, 2016.
15. K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, 15(3):147–204, 2005.
16. J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, W. Lasecki, K. Gajos, and N. Heffernan. Axis: Generating explanations at scale with learnersourcing and machine learning. In *ACM Learning at Scale*, 2016.
17. B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009.