Evaluating Crowdworkers as a Proxy for Online Learners in Video-Based Learning Contexts

DAN DAVIS, Delft University of Technology, the Netherlands
CLAUDIA HAUFF, Delft University of Technology, the Netherlands
GEERT-JAN HOUBEN, Delft University of Technology, the Netherlands

Crowdsourcing has emerged as an effective method of scaling-up tasks previously reserved for a small set of experts. Accordingly, researchers in the large-scale online learning space have begun to employ crowdworkers to conduct research about large-scale, open online learning. We here report results from a crowdsourcing study ($N = 135$) to evaluate the extent to which crowdworkers and MOOC learners behave comparably on lecture viewing and quiz tasks—the most utilized learning activities in MOOCs. This serves to (i) validate the assumption of previous research that crowdworkers are indeed reliable proxies of online learners and (ii) address the potential of employing crowdworkers as a means of online learning environment testing. Overall, we observe mixed results—in certain contexts (quiz performance and video watching behavior) crowdworkers appear to behave comparably to MOOC learners, and in other situations (interactions with in-video quizzes), their behaviors appear to be disparate. We conclude that future research should be cautious if employing crowdworkers to carry out learning tasks, as the two populations do not behave comparably on all learning-related activities.

CCS Concepts: • **Applied computing → Interactive learning environments**;

Additional Key Words and Phrases: Learning Analytics; MOOCs; Replication; Crowdwork

ACM Reference Format:
Dan Davis, Claudia Hauff, and Geert-Jan Houben. 2018. Evaluating Crowdworkers as a Proxy for Online Learners in Video-Based Learning Contexts. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 42 (November 2018), 16 pages. https://doi.org/10.1145/3274311

1 INTRODUCTION

Massive Open Online Courses (MOOCs) have opened the door to an entirely new discipline of large-scale learning analytics, or the study of learning behavior through the analysis of digital traces left by learners in large-scale online learning environments.

Creating a MOOC, however, is time and resource intensive—from the instructional design of the course to the production and recording of the lecture videos to the development of the course in the online platform, it is a substantial investment. It is also high-stakes due to its openness and scale—the entire internet-connected world has access. To address these factors we explore the extent to which online learning researchers and practitioners can employ crowdworkers to conduct learning analytics research and testing of online learning environments, namely MOOCs. Recent research in other disciplines such as psychology, information retrieval, and medicine has

Authors' addresses: Dan Davis, Delft University of Technology, Delft, the Netherlands, d.j.davis@tudelft.nl; Claudia Hauff, Delft University of Technology, Delft, the Netherlands, c.hauff@tudelft.nl; Geert-Jan Houben, Delft University of Technology, Delft, the Netherlands, g.j.p.m.houben@tudelft.nl.

2573-0142/2018/11-ART42 $15.00
https://doi.org/10.1145/3274311

done similar work and successfully leveraged crowdworking platforms (such as Crowdflower or Amazon MTurk) to streamline and optimize a wide variety of tasks, ranging from data labeling to user testing at scale [1, 2, 6, 17, 32]. Crowdworkers are a formidable population to which such testing may be delegated due to their massive scale in contrast to the scarcity of domain experts in a given task. Some have already seen its potential and begun to employ crowdsourcing platforms as a way to conduct large-scale online learning research—operating under the implicit assumption that the two populations (MOOC learners and crowdworkers) behave comparably. We here explore the extent to which the behavior of crowdworkers does indeed approximate that of MOOC learners and thus investigate the validity of this assumption because, before now, this has not been evaluated.

We operate under the hypothesis that crowdworkers could potentially aid in two key aspects of online learning: (i) researchers can conduct learning analytics experiments in low-stakes environments with reliable transfer/generalization to an actual learner population and (ii) universities or practitioners can test their educational resources in a low-stakes environment before rolling it out widely. For example, learning experiences and resources are often thoughtfully designed through an iterative process meant to shepherd the learner through a carefully designed pathway towards the formation of some new knowledge. However, research has shown that online learners do not always follow these prescribed learning paths and can sometimes undertake unpredictable trajectories through learning processes [10, 16, 27, 35, 47]. Crowdsourcing can here take some of the uncertainty out of this process and identify any undesirable behaviors or trends to be ameliorated in the iterative design process. While the latter aspect (testing educational resources) is more novel and unexplored, the former (conducting research) has been a growing trend in the learning analytics research literature [13, 25, 33]. This study will explore the extent to which findings from experiments run using crowdworkers transfer to real online learners by replicating the methods and analyses used in a selection of published works.

While not commonly reported on in the peer-reviewed research literature, the testing of courses before official launch is currently widely practiced using a variety of methods. For example, Coursera reportedly has an assemblage of 2,500 volunteers who beta test courses before they are widely released [1]. Duolingo makes the beta version of their course open to anyone who chooses to enroll—understanding that it is not the complete, final course and that they may experience technical issues [2]. Another method frequently used is to have a small team of internal employees or students manually inspect the quality of each course before launch. These approaches all differ from leveraging the crowd in that testers are explicitly asked to evaluate a course's technical function and content clarity, thus creating bias in the way they engage with the course. With crowdworkers, on the other hand, their experience through the course materials can be framed not as a testers, but as learners—with the objective of learning rather than testing—thus enabling more accurate analyses of learning behavior in the course environment.

To address this topic, we replicate a range of learning analytics experiments centered around video-based learning carried out in live MOOCs and assess how well the behavior patterns observed in those experiments can be observed in a crowdwork context. We focus specifically on video-based learning contexts, as this is the primary teaching mechanism in contemporary MOOCs. We arrive at the following guiding **R**esearch **Q**uestion:

RQ To what extent can the behavior of crowdworkers serve as a reliable proxy to that of MOOC learners?

To evaluate this question we conduct comparisons between the behavior of MOOC learners and crowdworkers in similar contexts across key online learning engagement measures. We evaluate

[1] https://bit.ly/2kR0I5P
[2] https://bit.ly/2J8G7S6

our research question based on three types of learning activities: standard video lectures, in-video quizzes, and post-video quizzes. Based on our exploration of these, we make the following contributions:

- The comparability of behavior between MOOC learners and crowdworkers is dependent on the type of learning activity being evaluated
- MOOC learners and crowdworkers engage comparably with traditional lecture videos.
- Crowdworkers and MOOC learners do not engage similarly with in-video quizzes.
- Crowdworkers perform comparably to MOOC learners on post-lecture knowledge assessment activities.

Furthermore, we contribute a set of design principles specifically for learning-based crowdsourcing tasks to elicit comparable behavior between crowdworkers and MOOC learners (in terms of quiz performance and video watching behavior). We emphasize that the findings and design principles presented below are derived from a specific context (learning-based video engagement and behavior), and that future research should critically evaluate the extent to which these findings transfer to other contexts beyond the one explored here. We also offer an open-source software (`VidQuiz`) tool for practitioners and researchers to use for in-video quiz activities.

2 RELATED WORK

In this section we first describe similar work from other fields that has shown the potential of crowdworkers to perform certain tasks as reliably as experts and domain-specific users. We then outline a set of studies which have employed crowdworkers in a learning context to gain insights about online learning.

2.1 Crowdsourcing

Defined as "the act of taking a job traditionally performed by a designated agent and outsourcing it to an undefined, generally large group of people..." [22] crowdsourcing harnesses the power of the (massive and online) crowd in order to complete a variety of tasks at an unprecedented speed and magnitude.

Crowdsourcing first emerged as an alternative to traditional lab experiments in the social sciences around 2010 [5]. A group of early adopters found that they were able to both validate and extend their findings from a small lab setting to the massive scale of the crowd [31]. Since then, researchers have been leveraging crowdwork platforms as a recruitment tool to conduct research at a rapid pace and relatively low cost.

Researchers from a wide variety of fields have begun to explore the possibilities afforded by crowdsourcing to expedite their data collection and evaluation [15]. Such experiments have found crowdworkers to be suitable for tasks of varying complexity as well—from simple survey tasks to complex labeling tasks previously carried out by experts [6]. Other research, such as that outlined in [12], explores methods for training crowdworkers to complete more complex tasks than they otherwise could using learning science principles and theory.

[1] conducted a study evaluating the reliability of crowdworkers to complete relevance assessment tasks. In this experiment, the authors measured the agreement between crowdworkers and expert assessors in determinig the relevance of a given document to a given topic. Coming from the information retrieval discipline, this information is used to compute effectiveness metrics in order to train document ranking algorithms. The results show that "[crowd]workers not only are accurate in assessing relevance but in some cases were as precise as the original expert assessors, if not more" [1]. A number of other experiments [2, 17, 23, 26] have been successfully carried out with similar results.

In another context, [6] conducted a study which compared the reliability of health record annotations created by crowdworkers to those created by experts. Results indicate that annotations (in this case on radiology records) made by crowdworkers are just as reliable and accurate as those made by experts.

Similar to the studies described above, [42] employed crowdworkers to conduct content analyses on textual data for use in psychological research. In this study, not only did the authors compare the crowdworker reliability to experts, they also compared it to previously published results and automated machine labeling. Results show that crowdworkers were equally reliable as both experts and published results and outperformed the computer software.

These studies discussed above are all outcome-centered—they have a task performed by experts (usually labeling) and then explore how to get the same labels from crowdworkers. In the present research, however, we place learners in the "expert" equivalent role and consider more behavioral/engagement metrics because learning is an abstract and longitudinal process, especially compared to the micro, concrete nature of traditional crowdwork tasks.

2.2 Crowdsourcing in Learning Research

Researchers in the online learning field have recently begun to apply crowdsourcing to learning topics [11]. [14] theorizes about the employment of crowdworkers as learners by considering "...a crowdworker as a learner in an atypical learning environment." In this study the authors outline some of the individual differences between learners and crowdworkers: (i) since the tasks are short in nature, crowdworkers must learn "on the fly," and (ii) crowdworkers face a very low chance of ever applying their lessons from a crowdsourcing task, which makes them less inclined to commit to learning.

Crowdworkers have been employed to evaluate the effect of: (i) various metacognitive prompting strategies for learners [13, 25, 33], (ii) different strategies for formative assessment & questioning [3, 50], (iii) learner feedback [34, 37, 49], (iv) cooperative learning environments or activities [7], (v) gamification and learning simulations [4, 9], and (vi) interactive multimedia learning activities [30, 43, 46]. Findings from each of these studies were meant to transfer/generalize to online learners but do not account for population differences, such as the fact that the participants were paid and online learners are not.

[8] conducted a study which combined learning and labeling in a crowdsourcing task with the dual purpose of measuring language learning and simulating experts in a task. The authors found that crowdworkers who were tasked with editing foreign-language video annotations were reliable at the editing task and also showed signs of language learning.

In the present study we explore the validity of the assumption underpinning these studies that results observed with crowdworkers are transferable to actual learning contexts.

3 METHOD

In this section we first describe the two main studies we replicate in the crowdsourcing context and then describe the procedure, materials, and measures employed.

3.1 Replicated Studies

In this study we replicate analyses primarily from two studies (Kovacs [29] & Kim et al. [24]) and supplement these results with comparisons to other results reported in the MOOC research canon on learner behavior. We selected these two primary studies for replication for three key reasons: (i) they are about MOOC learners' engagement with videos, which serve as the principal means of instruction in most online learning environments [18, 24, 38–40, 45, 46], (ii) they present numerous quantitative, empirical results with enough methodological detail to inform and enable a

thorough replication, and (iii) they cover short-term behaviors (in contrast to longitudinal measures such as long-term knowledge retention), as crowdworkers cannot be expected to engage in such time-intensive tasks. From our search we found that fulfilling these criteria is very uncommon, and Kovacs [29] & Kim et al. [24] emerged as the most suitable candidates for replication.

3.1.1 In-Video Quizzes. In [29], the authors analyzed data from a MOOC containing 92 videos with in-video quizzes (ungraded quiz questions that are embedded in a video and appear as an overlay at certain points in the video; videos contained either one or two in-video quiz questions). They analyze learner engagement (measured using clickstream events of their interactions) with these in-video quizzes in great depth and uncover common trends that emerge over the four-month span of the course. The key findings are as follows: (i) there are peaks in in-video seeking activity surrounding in-video quizzes, (ii) learners most frequently seek backwards from in-video quizzes, and (iii) most learners answer in-video quiz questions correctly on their first attempt. The in-video quiz questions were ungraded and not required for learners to pass the course.

3.1.2 Interaction Peaks. In their study of 862 MOOC videos accounting for over 39 million interaction events (e.g., play, pause, and seek video), the authors in [24] identified student activity patterns that account for high concentrations of video engagement events (peaks) at distinct moments in videos indicating time-specific interest. The two video elements that most commonly cause peaks in learners' video interactions are (i) starting a new topic in the form of a visual transition (e.g., from a slide/text to a talking head) and (ii) important non-visual explanations (e.g., explanations about key topics without visual support). They found that these patterns generalize not only across videos but even courses on varying topics as well.

We compare the findings of these studies to our own results from crowd workers led through a learning activity by directly comparing quantitative measures of behavior (through log traces) as well as qualitative comparisons of graphical representations of the data on a case-by-case basis.

3.2 Procedure & Materials

Participants were recruited using the CrowdFlower[3] platform and provided a Token and a link to the `VidQuiz` web application (described in detail in Section 3.4). We elected to employ CrowdFlower as it has been found to attract a highly diverse workforce in terms of demographics through its partnerships with a wide variety of workforce providers [44]. This is an important characteristic for the present study, as we are chiefly concerned with MOOCs, which are open to and used by people from all over the world. Furthermore, CrowdFlower offers the ability to direct workers to custom, external environments and return with a token of completion. Crowdflower is also recognized for its robust, built-in quality control mechanisms [44]. Participants were instructed to read the task rules and directions before clicking the link and beginning the task. The only demographic requirement we imposed was that participants must be proficient in English. We did not place any geographic restrictions on participants in the experiment as there are no geographic restrictions on MOOC enrollment or participation. Participants represent 37 different nationalities (including the most common countries for MOOC learner demographics: United States, India, and China); the top three most represented countries in our study were Venezuela, Serbia, and Egypt.

The web pages in our `VidQuiz` web application appeared in the following order:

1. Task Introduction & Token Entry
2. Video 1 (topic: Solar Energy, duration: 7m40s)
   - Original MOOC lecture Video (for interaction peaks)
3. Quiz 1 (Solar Energy)

[3] www.crowdflower.com

   - Answer 4 multiple choice quiz questions about video 1
4. Video 2 (topic: Design Prototyping, duration: 6m30s)
   - Original MOOC lecture video with 2 multiple choice in-video quiz questions
5. Quiz 2 (Design Prototyping)
   - Answer 3 multiple choice quiz questions about video 2
6. Exit Token Assignment

The participants would then submit the exit token to CrowdFlower to receive their payment after taking a single-question post-task survey.

We selected these materials because they were designed to be accessible to all (no course pre-requisites) and are the ideal/recommended length of MOOC lecture videos [18, 36]. The videos were selected for specific reasons; in the two studies being replicated here, results are from a large sample of videos which, when considered together, contain a highly diverse range of characteristics (such as visual transitions). The video selected for the interaction peak analysis was chosen because it contains (i) a key non-visual explanation, (ii) visual explanations accompanied by visual transitions, (iii) a clear, engaging speaker, and (iv) entry-level content accessible to those not taking the full course, thus making it well representative of the types of videos used widely in MOOCs. Due to the finding by [29] that the presence of in-video quizzes conflates the normal interaction peak patterns, we selected a lecture video for the in-video quiz portion & analyses without a slide-show-style presentation.

Navigating to the previous page of the web application was only possible during the two quizzes, where participants were welcomed to go back and search for the answer in the lecture video. All quiz questions and videos were sourced directly from existing MOOCs.

Although they were repeatedly reminded and encouraged to answer each question to the best of their ability, we did not require participants to answer questions correctly to advance. Even though many crowdworking tasks are chiefly concerned with response accuracy, that stipulation is not transferable to a learning context, as the purpose of a learning assessment is to evaluate each individual's knowledge state regardless of its standing. In the present case, instead of using pre-task qualifications, we filter quality participants by requiring them to remain in the active browser tab and watch at least 90% of the video. Each crowdworker was only allowed to complete the task one time. Upon successful completion of the task, each participant was awarded $1 USD.

In order to ensure that workers exhibit reasonable effort to learn the material (instead of simply opening a new tab/window and searching the web for the answers to the quiz questions), we took the following measures to promote honest behavior for the task by (i) asking quiz questions that can be readily answered from the lecture video but at the same time are not common knowledge and (ii) telling participants that they may not leave the active window or change tabs to search the web for the answer—and that if we detect that they do indeed leave the tab more than a certain number of times (4)[4], they will be disqualified from the experiment. We enforced this policy by using JavaScript to monitor whether or not the browser tab with the experiment was active or not. If a participant violated the active window rule, they were disqualified from the experiment and could not advance through the system. Of all crowdworkers to begin the task, 33% fully adhered to the rules and thus successfully completed the task.

3.3 Measures

The following events are all logged through the `VidQuiz` web application along with their respective time stamps and used in the ensuing analyses. These measures are all defined and operationalized in the same manner as those in the studies being replicated.

[4]We did not restrict participants to never leaving the tab at all because that would have been too restrictive. By giving a budget of 4 tab exits, it gives participants reasonable flexibility and ensures reliable results.

- Navigation events
  - `Begin User`: fired when a user visits the first page of the experiment
  - `Page Loaded`: fired when a user visits any page of the web application
  - `Quiz Question Response`: each response from the quizzes following the lecture videos
- Tab visibility:
  - `Tab Hidden`: fired when a user leaves the browser tab housing the experiment. On pages with videos, this event also triggers the video to pause.
  - `Tab Visible`: fired when a user returns to the browser tab with the experiment
- Video Interactions
  - `Play`: fired when the video begins playing
  - `Pause`: fired when the video is paused
  - `Seek`: fired when the user manually seeks through the video time line. These events also contain the origin (seekFrom) and destination (seekTo) of the seek event.
  - `Rate Change`: fired when a user changes the playback speed of the video
  - `In-Video Quiz Question Delivery`: fired each time in-video quiz questions appear
  - `In-Video Quiz Question Response`: recording the user's responses to the question

3.4 `VidQuiz` Overview

We developed `VidQuiz`[5] (shown in Figure 1) specifically for the present experiment upon realizing the lack of freely-available in-video quizzing resources. In this section we provide a brief description of the system architecture followed by a design rationale for the web application's user interface.

3.4.1 System Architecture. The `VidQuiz` web application follows a RESTful client-server architecture to store the logs of users' behavior. For the exhaustive list of events that are logged and stored, refer back to Section 3.3. The system front end has two main functionalities: (i) track and persist learner activity data to the server and (ii) display the user interfaces. To enable real-time storage and retrieval of user activity data from the back-end, we implemented an HTTPS server in Node.js and persisted the tracked events to a MongoDB database.

3.4.2 User Interface Design. We developed the front-end interface for `VidQuiz` by combining a trio of open-source libraries and plugins. To display the lecture videos we utilized `Video.js`[6], an open-source JavaScript library for customizing HTML video players. To display the quiz questions we utilized `Survey.js`[7], an open-source JavaScript survey library. The technique for delivering in-video quiz questions required the combination of the `Video.js` player, `Survey.js` question interface, and a `Video.js` plugin named `Videojs-markers`[8] to show users the locations of in-video quiz questions along the video player time line (yellow bars shown in Figure 1a).

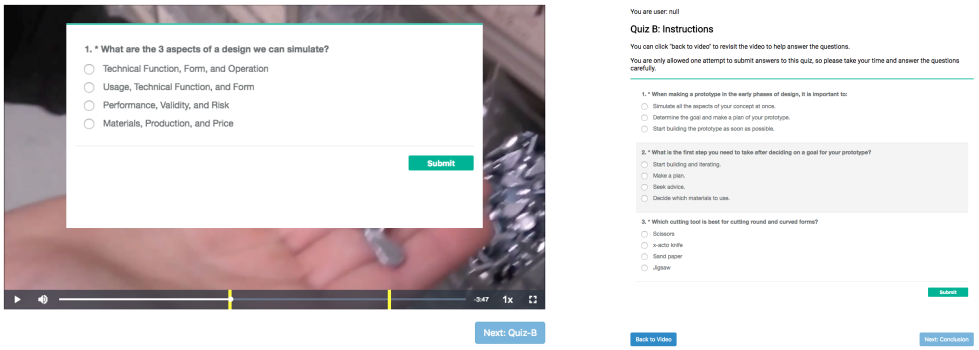
4 FINDINGS

In this section we first present findings from a pilot study. We then present results from the main study, comparing MOOC learners and crowd workers in terms of general behavior, in-video quiz engagement patterns, video interaction peaks, and quiz scores.

4.1 Pilot Study Findings

We conducted a pilot study ($N = 86$) to (i) evaluate the technical function of the `VidQuiz` system and (ii) conduct preliminary analyses of participants' behavior within the system by replicating analyses from [29] and [24] to better understand user behavior in the system.

[5]Open-source available at [withheld for blind review]
[6]https://videojs.com/
[7]https://surveyjs.io/
[8]http://sampingchuang.com/videojs-markers



(a) VidQuiz Video Player with an in-video quiz question (b) Post-Video Quiz

Fig. 1. Screenshots from the VidQuiz interface.

Consistent with [29], we observe a trend indicating that once users interact with the first in-video quiz question, they begin to understand the system more and are more inclined to seek ahead to the second one. To help accelerate this learning curve, in the main study we inserted a "Guide" page before the video lecture with in-video quizzes. This page gave brief instructions on what to expect and how to interact with the in-video quiz interface.

To address the crowdworkers' poor performance on quizzes in the pilot study (31% correct overall), we speculated that learners expected an opportunity for a second attempt at the question should they answer incorrectly. We therefore provided reminders to participants that they are only allowed one attempt per question. In the same vein, we also suspected that the poor performance on the quiz questions was a result of participants not realizing that once an in-video quiz question appears, they can close it, seek back in the video, find the answer, and return to the question to answer it. Likewise for post-video quizzes—users can leave the quiz page and go back to the video to search for the answer. To address this we added more cues reminding participants of this affordance and also added a point about this on the Guide page mentioned above.

4.2 Crowdworker Behavior in VidQuiz

For the main study ($N = 409$), before analyzing the participants' behavior in comparison to MOOC learners according to Kovacs [29] and Kim et al. [24], we first explored our research question through some general behavior metrics to get a sense of how crowdworkers engaged with the experiment as a whole and if these behaviors align with results reported in the MOOC research literature.

Figure 2a shows the total time on task (measured from the moment they open the first page of the task to the moment they unlock their token of completion) of those who completed the task (completers, $N = 135$) and those who did not (non-completers, $N = 274$). We find that the median time spent for completers is 21 minutes and that of non-completers is 11 minutes—that is to say they typically spent 11 minutes on the task before either dropping out on their own accord or having the system cut them off for violating the task rules.

This path of attrition is broken down in Figure 2b, where we show the total number of times each page in the system/task was loaded by both completers and non-completers. We observe a steep decline for non-completers during/after the first video lecture (Solar Energy), presumably because these participants realized that they were not able to skip past the video and actually had to watch

it in its entirety; this acted as a deterrent to crowdworkers who were not willing to commit to such a task. We observe another steep decline during/after the second video lecture (Prototyping), and some participants even drop out at Quiz-B, just one click away from task completion. This is equivalent to the established slope of attrition by MOOC learners replicated and confirmed across a number of previous studies [21, 28]—albeit on a timescale of weeks, whereas the crowdworkers' attrition happens in a matter of minutes.

(a) Total time spent on task (b) Attrition: loads per page (c) Completers' self-reported activity during lecture videos

Fig. 2. Aggregated overview of participant behavior during the task

One concern going into the main study was the question of whether or not crowdworkers would find shortcuts that allow them to unlock the token of completion without thoughtfully engaging in each activity. Based on Figures 2a and 2c, we see that this was not the case—as shown in Figure 2a, those who completed the experiment spent a median total time of 21 minutes on the task, and 87% of them self-reported that they watched the lecture videos attentively in a required post-task survey question (cf Figure 2c), thus indicating that tasks can be structured in a way that elicits users' undivided attention.

Just as prior work has found that certificate-earning MOOC learners undertake a far more linear path than those who do not go on to finish the course [10], we likewise observe that those who complete the task navigate the `VidQuiz` system in a more linear fashion than non-completers; we measure this by the average number of page loads per participant, as more page loads indicates more deviation (so a value of 1.0 would indicate a participant only opened each page one time and always navigated to the following one, never tracking back): 1.3 vs. 4.5 respectively (the difference is statistically significant as determined by a one-way ANOVA $F = 65.7$, $p < 0.0001$).

4.3 In-Video Quizzes

In our replication of the study by Kovacs [29] about MOOC learners' engagement with in-video quizzes, we evaluated our research question by reporting quantitative comparisons between our study and the original and then presenting qualitative comparisons to the data visualizations provided by the authors.

Table 1 presents quantitative comparisons between MOOC learners and crowdworkers while engaging with in-video quiz videos across eight key measures. When comparing the two populations one important difference is that responding to in-video quizzes in `VidQuiz` is required to complete the task, whereas in-video quiz questions in the context of Kim et al. [24] were optional.

Highlighting some of the discrepancies from Table 1, we consider the sizable difference in performance on the in-video quiz questions (answering correctly), with learners in Kovacs [29]

faring 18 percentage points better than crowdworkers. The largest discrepancy in Table 1 comes in the way the two populations engage *after* in-video quizzes. When faced with an in-video quiz, MOOC learners' in Kovacs [29] are 55x more likely to perform a backwards seek and 4.1x more likely to perform forward seek than they are at any other point in the video. This trend is reversed for crowdworkers, who are far more likely to seek forward than backwards after being presented with an in-video quiz. This trend is sensible given the difference in performance between the two groups—MOOC learners learn the strategy (over the span of 90+ videos in the course) of seeking backwards to find the correct answer to the question from the video (and thus answer correctly 76% of the time) whereas crowdworkers either choose not to adopt this strategy or do not consider it and make a less-informed attempt at answering the question.

Table 1. IVQ (In-Video Quiz): Quantitative comparisons between the crowdworkers observed in the current study versus MOOC learners described by Kovacs [29]

	MOOC Learners [29]	Crowdworkers
Start Video → Finish Video	68%	83%
Start Video → Answer IVQ (Completers)	79%	100%
Start Video → Answer IVQ (Non-completers)	72%	62%
Answer IVQ Correctly	76%	54%
Avg. Length of Seek Event	31s	29s
% of Chains Skipping Ahead	56%	70%
Backward Seek Rate from IVQ	55x	7.7x
Forward Seek Rate from IVQ	4.1x	49.3x

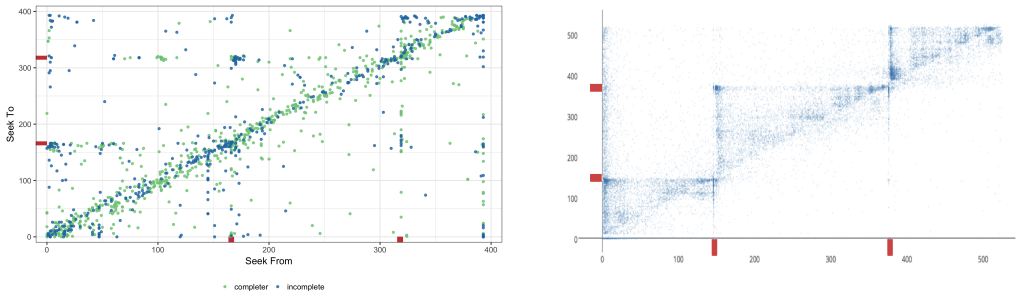
Figure 3 shows the seek event patterns of crowdworkers from the current study and MOOC learners from Kovacs [29]. Figure 3a indicates that non-completers are more prone to seeking directly to the end of the video as well as in-video quizzes. This is in stark contrast to the behavior of completers, who advance through the video in a linear fashion. This behavior indicates that non-completers tried to game the system and advance through to the next page without actually watching the video (and, consequently, went on to exit the experiment).

From Figure 3 we see that the most similar characteristic between the two populations is the concentration of events around in-video quizzes. We also see evidence that, even though crowdworkers rarely seek backwards, when they do so it is usually from a quiz. The most prominent difference between the two graphs in Figure 3 is the linear fashion through which crowdworkers log seek events; compare this to the more stepped pattern of MOOC learners. These qualitative comparisons between visualizations are primarily limited by the sample size represented in each. There are data from 61,453 learners represented in the [29] study, accounting for over 6.4 million seek events distributed across 92 videos. The present study includes 409 participants in total who account for 1,686 seek events distributed across 2 videos.

Overall, although there are a number of similarities such as the average length of seek events and high concentration of seek events around in-video quizzes, we find that the two populations are not similar across enough behavioral metrics to be considered comparable.

4.4 Interaction Peaks

We next explored the similarity between MOOC learners and crowd workers with respect to video interaction peaks. In [24] the authors found that 61% of identified peaks are associated with a visual transition in the lecture video. As illustrated in Figure 4a, in our replication we observe evidence



(a) Crowdworkers

(b) MOOC Learners, from [29]

Fig. 3. Comparing the results of the in-video quiz seeking behavior analysis to those found in [29]. Coordinates (x,y) represent a seek from x to y. Red markers on each axis show the location of in-video quiz questions.

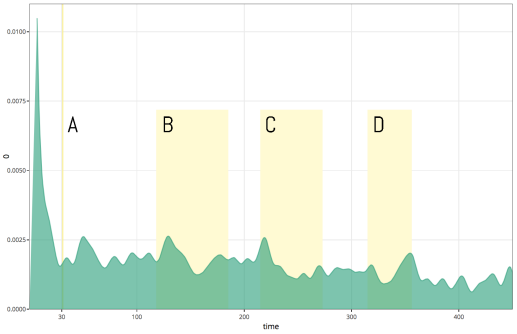
to support this finding: two of the three key visual transitions (labels B and C, where the video switches from the lecturer to a graphic/slide and back) in the video lecture led to interaction peaks shortly after the transition occurs.

The vertical yellow bar in Figure 4a (label A) at 30s shows where the answer to a post-video quiz question was said (not shown) by the speaker. The highlighted regions B and C each contain the (visual) answer to quiz questions. For those two, we note that interaction peaks occur shortly after the visual appears, and for the final transition (highlighted region D), we see distinct interaction peaks at the beginning and end of the visual—all consistent with findings from Kim et al. [24]. In this graph we employed the same "bin, summarise, and smooth" plotting method outlined in [24] and adopted from [48].

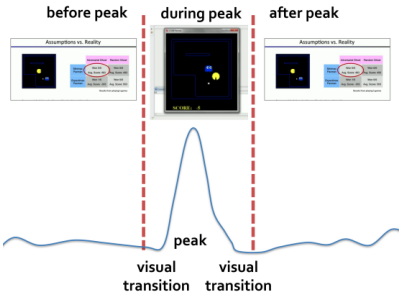
The authors in [24] found that the mean peak height was 7.7% (std=10.4) of the maximum (caused by videos playing automatically upon page load). Among those to have completed the study, we found the maximum interaction peak to have a density of 0.026 (y-axis in Figure 4), and we found the three highlighted interaction peaks to have a respective density of 0.0028 (10.8% of the maximum), 0.0023 (8.8%), and 0.0015 (5.8%)—each falling in the expected range of values compared to [24]. [24] also found that, on average, there were 3.7 peaks per video in their dataset; we likewise observed 4 peaks in ours for this lecture video.

The authors in [24] found that the median peak width for all video types was 9 seconds. In line with their finding that lecture video types have a more continuous flow and lead to wider, less sharp interaction peaks (compared to tutorial videos, for instance), we found that the peaks in highlighted regions B, C, and D in Figure 4a have a width of 25 seconds, 20 seconds, and 15 seconds respectively.

[24] also reports that a key cause of interaction peaks (accounting for 39% of the total) is non-visual explanations (verbal instructions with semantic importance). As shown in Figure 4a, an example of this occurs at the 30 second mark in the video. It is at this moment when the lecturer verbalizes the definition of energy—which appeared on the post-video quiz—without any visual support (the answers to the other quiz questions also had visual/textual support). Even though this peak is not as prominent as those in the highlighted regions with visual transitions (which is consistent with the finding that "peaks were taller and larger in size when they had visual transitions nearby" [24]), this small peak does indeed indicate that learners express a time-specific interest in this section of the video to answer the quiz question.



(a) Interaction peaks by crowd workers. Spikes show concentrations of play/pause events. The highlight at point A indicates the point where the speaker verbalized a key concept. Highlighted regions B, C, and D show the times with key concepts shown via text.



(b) From [24]: The structure of an interaction peak (for MOOC learners) surrounded by visual transitions before and after.

Fig. 4. Comparing the results of the interaction peak analysis to those found in [24].

Overall, we observe comparable trends between the behavior of crowdworkers in the present study and MOOC learners reported by [24]. This indicates that, in a highly controlled and designed experimental environment, crowdworkers will exhibit similar behavior to MOOC learners in the way they interact with lecture videos.

4.5 Crowdworker Quiz Scores

To explore the third aspect of our research question, we next compared MOOC learners vs. crowdworkers in terms of their performance on standard post-lecture quizzes.

Just like the in-video quiz questions, these questions were taken from a MOOC and could readily be answered by only using content from the video lecture. On the first quiz (about solar energy), crowdworkers who completed the task answered 61% of questions correctly. And for the second quiz (about design prototyping): 56% correct.

We found that participants who completed the study spent an average of 2.5 minutes on the quizzes. Considering that the quizzes only consisted of three or four multiple choice questions, we consider 2.5 minutes adequate time to give the necessary amount of thought to solving the questions.

To compare the crowdworkers quiz scores to a baseline standard of MOOC learner quiz scores, we reference the results reported in [28]. The authors found that MOOC learners scored an average of 63% on quizzes and 57% on a final exam. This rate is also consistent to the results reported in [19] where MOOC learners earned an average score of 65% on the course final exam. Compared to the average scores of crowdworkers that we report above (61%), we find the two populations to be highly similar in this regard.

5 DISCUSSION

Whereas motivation and self-regulation have emerged as the most important factors leading to successful MOOC learning outcomes [27], a willingness to follow rules and pay attention to the task are the decisive factors leading to the successful completion of a video-based learning crowdwork task.

In designing the study and task environment, the main challenge to consider was accounting for the vastly different motivation profiles of MOOC learners vs. crowdworkers. MOOC learners do not earn any money by completing a course, but they do earn a certificate of accomplishment as well as any knowledge gained from the experience. Crowdworkers, on the other hand, are solely concerned with earning money and have no concern for learning or career advancement.

However, MOOC learners are unlike traditional university students who have paid considerable money and made a serious commitment to a degree program. It is not uncommon for a MOOC learner to enroll in a course, watch one or two video lectures, and then never return. If that learner gained what he or she was hoping to gain from those videos—even though they would have left the course with a grade of zero—then that is seen as a success. What this means in the crowdsourcing context is that the task must be designed in a way to allow for such whimsical behavior—where participants are free to act on their own accord, but if they do not follow the rules clearly outlined (with frequent reminders) for the task, then they will not be eligible for the payout.

From the present study we therefore offer the following principles for designing video-based learning tasks for crowdworkers: (i) limit behavior measurements to short-term outcomes (as opposed to longitudinal measures), (ii) select educational resources (such as readings or lecture videos) that do not require and pre-requisite knowledge, (iii) only use assessment questions that can be answered from the provided resources, (iv) provide frequent reminders of the task rules and platform affordances, and (v) restrict participants' tab browsing activity to keep them focused on the task while still allowing some flexibility.

We also note an important consideration for future research in this area: confirmatory null hypothesis testing is not designed to make claims of similarity between two samples. Rather, such statistical tests are solely intended to conclude significant differences. We emphasize this consideration going forward, as future research should be framed in a way to either conclude that the two populations (MOOC learners and crowdworkers) behave differently, or that no significant difference could be observed. An alternative solution is the application of equivalence tests, which evaluate whether confidence intervals fall within a set of defined equivalence bounds [20]. These could also serve to curtail misinterpretations of p-values from null hypothesis tests—as values far greater than the alpha do not indicate a greater degree of insignificance than those closer to (and still greater than) it. In the present study, we did not use such testing to compare the two populations, as insufficient data was provided in the replicated works.

In conclusion, we here report on a study which employs crowd workers to evaluate the extent to which they can serve as reliable proxies for MOOC learners for the purposes of both learning analytics research and testing. Specifically, we find that (i) the crowd workers behave comparably to MOOC learners in their engagement with standard lecture videos, (ii) the two populations do not engage similarly with in-video quiz question materials, and (iii) the two populations perform comparably on assessment/quiz activities. We also contribute `VidQuiz`, the open-source in-video quizzing software created to carry out the experiment.

With two of the three video-based learning activity/behavior types leading to comparable results between MOOC learners and crowdworkers, we emphasize that the current results are not absolute in terms of the generalizability between the two populations. Learning environments are complex ecosystems with myriad factors at play, and careful consideration must be taken with regard to the expected transfer/generalizability of the results observed in the present study to other contexts. Accordingly, future research should not indiscriminately follow the guidelines offered here. Rather, due to the prominence of heterogeneous treatment effects in the learning sciences (how certain interventions or strategies affect certain types of learners differently), one must carefully consider the context in which the learning is taking place and being evaluated, as this often has a substantial effect on behavior and measured outcomes [41]. Even in adaptive/personalized learning

environments or platforms, claims and estimations of what a given learner needs at a given moment are limited to that context, and the same models could not be readily transferred to a new course, topic, or platform without thorough critical evaluation. As evidenced by the present study, not all types of learning activities translate to the crowdwork context equally, so more research must be done in working towards a complete understanding of which learning activities can be reliably delegated to the crowd for testing and research. In doing so, researchers can continue to find ways to leverage the crowd to explore new approaches to innovating online learning environments and teaching at scale.

REFERENCES

[1] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information processing & management* 48, 6 (2012), 1053–1066.

[2] Omar Alonso, Daniel E Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, Vol. 42. ACM, 9–15.

[3] Yigal Attali. 2015. Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers & Education* 86 (2015), 260–267.

[4] Yigal Attali and Meirav Arieli-Attali. 2015. Gamification in assessment: Do points affect test performance? *Computers & Education* 83 (2015), 57–63.

[5] John Bohannon. 2011. Social science for pennies. (2011).

[6] Anne Cocos, Ting Qian, Chris Callison-Burch, and Aaron J Masino. 2017. Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. *Journal of biomedical informatics* 69 (2017), 86–92.

[7] Derrick Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A. Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *CSCW '15*. 1139–1152.

[8] Gabriel Culbertson, Solace Shen, Erik Andersen, and Malte Jung. 2017. Have your Cake and Eat it Too: Foreign Language Learning with a Crowdsourced Video Captioning System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 286–296.

[9] Maria Cutumisu and Daniel L Schwartz. 2016. Choosing versus Receiving Feedback: The Impact of Feedback Valence on Learning in an Assessment Game. In *EDM '16*. 341–346.

[10] Dan Davis, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2016. Gauging MOOC Learners' Adherence to the Designed Learning Path. In *Proceedings of the 9th International Conference on Educational Data Mining*. 54–61.

[11] Dan Davis, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2018. Activating learning at scale: A review of innovations in online learning strategies. *Computers & Education* 125 (2018), 327 – 344.

[12] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2623–2634.

[13] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 105–114.

[14] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. 2015. Training workers for improving performance in crowdsourcing microtasks. In *Design for Teaching and Learning in a Networked World*. Springer, 100–114.

[15] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. 2017. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, 6–26.

[16] Chase Geigle and ChengXiang Zhai. 2017. Modeling Student Behavior With Two-Layer Hidden Markov Models. *Journal of Educational Data Mining* 9, 1 (2017), 1–24.

[17] Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*. Association for Computational Linguistics, 172–179.

[18] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 41–50.

[19] Sherif Halawa, Daniel Greene, and John Mitchell. 2014. Dropout prediction in MOOCs using learner activity features. In *Experiences and best practices in and around MOOCs*, Vol. 7. 3–12.

[20] Walter W Hauck and Sharon Anderson. 1984. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics* 12, 1 (1984), 83–91.

[21] Andrew Dean Ho, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. 2014. *HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013.* Technical Report. Harvard University and Massachusetts Institute of Technology.
[22] Jeff Howe. 2008. *Crowdsourcing: How the power of the crowd is driving the future of business.* Random House.
[23] Gabriella Kazai and Natasa Milic-Frayling. 2009. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation*. 21–22.
[24] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. 2014. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 31–40.
[25] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1375–1386.
[26] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.
[27] René F Kizilcec, Mar Pérez-Sanagustín, and Jorge J Maldonado. 2017. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & education* 104 (2017), 18–33.
[28] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale*. ACM, 111–120.
[29] Geza Kovacs. 2016. Effects of in-video quizzes on MOOC lecture viewing. In *Proceedings of the third (2016) ACM conference on Learning@ Scale*. ACM, 31–40.
[30] Bum Chul Kwon and Bongshin Lee. 2016. A Comparative Evaluation on Online Learning Approaches using Parallel Coordinate Visualization. In *CHI '16*. ACM, 993–997.
[31] Chappell Lawson, Gabriel S Lenz, Andy Baker, and Michael Myers. 2010. Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics* 62, 4 (2010), 561–593.
[32] Xiao Ma, Megan Cackett, Leslie Park, Eric Chien, and Mor Naaman. 2018. Web-Based VR Experiments Powered by the Crowd. In *Proceedings of the 2018 World Wide Web Conference*. 33–43. https://doi.org/10.1145/3178876.3186034
[33] Jaclyn K Maass and Philip I Pavlik Jr. 2016. Modeling the Influence of Format and Depth during Effortful Retrieval Practice. In *EDM '16*. 143–150.
[34] Thi Thao Duyen T Nguyen, Thomas Garncarz, Felicia Ng, Laura A Dabbish, and Steven P Dow. 2017. Fruitful Feedback: Positive affective language and source anonymity improve critique reception and work outcomes. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1024–1034.
[35] Jihyun Park, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, and Mark Warschauer. 2017. Detecting changes in student behavior from clickstream data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 21–30.
[36] Oleksandra Poquet, Lisa Lim, Negin Mirriahi, and Shane Dawson. 2018. Video and learning: a systematic review (2007–2017). In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, 151–160.
[37] Anna N Rafferty, Rachel A Jansen, and Thomas L Griffiths. 2016. Using Inverse Planning for Personalized Feedback. In *EDM '16*. 472–477.
[38] Daniel T Seaton, Sergiy Nesterko, Tommy Mullaney, Justin Reich, Andrew Ho, and Isaac Chuang. 2014. Characterizing video use in the catalogue of MITx MOOCs. In *European MOOC Stakeholders Summit, Lausanne*. 140–146.
[39] Abdulhadi Shoufan. 2018. Estimating the cognitive value of YouTube's educational videos: A learning analytics approach. *Computers in Human Behavior* -, - (2018), –. https://doi.org/10.1016/j.chb.2018.03.036
[40] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. In *2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*. 1–12.
[41] David J Stanley and Jeffrey R Spence. 2014. Expectations for replications: Are yours realistic? *Perspectives on Psychological Science* 9, 3 (2014), 305–318.
[42] Jennifer Tosti-Kharas and Caryn Conley. 2016. Coding psychological constructs in text using Mechanical Turk: A reliable, accurate, and efficient alternative. *Frontiers in psychology* 7 (2016), 741.
[43] Selen Türkay. 2016. The effects of whiteboard animations on retention and subjective experiences when learning advanced physics topics. *Computers & Education* 98 (2016), 102–114.
[44] Donna Vakharia and Matthew Lease. 2015. Beyond Mechanical Turk: An analysis of paid crowd work platforms. In *Proceedings of the iConference*. 1–17.
[45] Frans Van der Sluis, Jasper Ginn, and Tim Van der Zee. 2016. Explaining Student Behavior at Scale: The influence of video complexity on student dwelling time. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 51–60.

[46] Tim Van der Zee, Wilfried Admiraal, Fred Paas, Nadira Saab, and Bas Giesbers. 2017. Effects of subtitles, complexity, and language proficiency on learning from online education videos. *Journal of Media Psychology: Theories, Methods, and Applications* 29, 1 (2017), 18.

[47] Miaomiao Wen and Carolyn Penstein Rosé. 2014. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 1983–1986.

[48] Hadley Wickham. 2013. *Bin-summarise-smooth: a framework for visualising large data.* Technical Report. had. co. nz, Tech. Rep.

[49] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *L@S '16*. 379–388. http://dl.acm.org/citation.cfm?id=2876042

[50] Joseph Jay Williams, Tania Lombrozo, Anne Hsu, Bernd Huber, and Juho Kim. 2016. Revising Learner Misconceptions Without Feedback: Prompting for Reflection on Anomalies. In *CHI '16*. ACM, 470–474.