Can Crowdsourcing Platforms Be Useful for Educational Research?

Karen D. Wang
kdwang@stanford.edu
Stanford University
Stanford, California, USA

Zhongzhou Chen
zhongzhou.chen@ucf.edu
University of Central Florida
Orlando, Florida, USA

Carl Wieman
cwieman@stanford.edu
Stanford University
Stanford, California, USA

ABSTRACT

A growing number of social science researchers, including educational researchers, have turned to online crowdsourcing platforms such as Prolific and MTurk for their experiments. However, there is a lack of research investigating the quality of data generated by online subjects and how they compare with traditional subject pools of college students in studies that involve cognitively demanding tasks. Using an interactive problem-solving task embedded in an educational simulation, we compare the task engagement and performance based on the interaction log data of college students recruited from Prolific to those from an introductory physics course. Results show that Prolific participants performed on par with participants from the physics class in obtaining the correct solutions. Furthermore, the physics course students who submitted incorrect answers were more likely than Prolific participants to make rushed cursory attempts to solve the problem. These results suggest that with thoughtful study design and advanced learning analytics and data mining techniques, crowdsourcing platforms can be a viable tool for conducting research on teaching and learning in higher education.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**; • **Information systems** → **Crowdsourcing**; **Data stream mining**.

KEYWORDS

crowdsourcing research, online experiments, postsecondary STEM education, log data, problem solving

ACM Reference Format:
Karen D. Wang, Zhongzhou Chen, and Carl Wieman. 2024. Can Crowdsourcing Platforms Be Useful for Educational Research?. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3636555.3636897

1 INTRODUCTION

Online crowdsourcing platforms have become increasingly popular for academic research over the past few years, especially since the Covid-19 pandemic limited researchers' ability to implement in-person lab studies. Despite the convenience and increasing popularity of online platforms like Amazon Mechanical Turk (MTurk) and Prolific, there remain questions about the quality of data collected through these platforms. Failure to address these questions is particularly problematic for education researchers, as inadequate or biased data would compromise the validity of the inferences made about students' characteristics or competencies and increase the risk of drawing false conclusions about the efficacy of specific instructional and intervention materials.

This study investigates whether crowdsourcing platforms like Prolific can serve as a viable data collection platform for educational research, especially for studies that involve cognitively demanding tasks designed to capture college students' higher-order competencies. Central to our inquiry is the use of an interactive problem-solving task embedded in an educational simulation, which records detailed log data on user interactions. In our previous research, we have identified log data-based behavioral features that correspond to specific problem-solving practices. The next phase of our research is focused on evaluating the feasibility and validity of automating the assessment of students' problem-solving practices at scale. Achieving this goal requires gathering a large dataset of students from diverse backgrounds working on the task, which brings into focus the utility of crowdsourcing platforms for such data collection efforts. Given this context, the goal of the current study is to conduct a comparative analysis of problem-solving performance between two diverse groups of participants: college students from an introductory physics course at a large university and US-based STEM undergraduates recruited through Prolific. Specifically, we aim to address the following research questions:

- RQ1: To what extent are the knowledge background and behavioral engagement levels comparable between Prolific participants and college students from a physics course?
- RQ2: To what extent are the problem-solving solution qualities comparable between the two samples?
- RQ3: To what extent are the problem-solving processes as captured by log data comparable between the two samples?

2 BACKGROUND

Prolific (www.prolific.com) is a web platform designed to facilitate online research by connecting researchers with potential participants [37]. Key features of the Prolific platform include 1) integrated recruitment, participation, and compensation processes; 2) pre-screeners capturing various information of participants (e.g., demographics, geographic, education and occupation) that allow researchers to obtain a sample with specific characteristics; 3) clear guidelines on compensation and specific criteria for approving/rejecting participant submissions.

In the field of education research, the platform has been used to collect data on the mental well-being of university students [12],

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636897

parents' descriptions of children's physical activity constraints during a pandemic [44], undergraduate and Master's students' acceptance of digital learning [51], and how adult learners assess their knowledge in the context of reading science topics [58]. These studies illustrate how Prolific can serve as an alternative to traditional data collection venues such as convenience sampling within a class or recruitment via community listservs.

2.1 Existing research on the advantages and limitations of online research platforms

There are several compelling advantages to the adoption of online crowdsourcing platforms for conducting research, including the ability to collect large amounts of data in a short period of time and the potential to obtain a more diverse sample than the traditional sampling pools of college students enrolled in specific courses [13]. Furthermore, previous research has generally found that the Prolific platform was at least comparable if not superior to MTurk [18, 57]. For instance, Peer et al. [42] reported that Prolific participants performed better on measures of attention, comprehension, and honesty than participants from MTurk and other platforms.

On the other hand, a number of studies have drawn attention to the potential downsides of collecting research data from online participants, such as inattentiveness, high attrition, and lack of sufficient knowledge in a specific domain. Chandler et al. [6] found that online participants were often engaged in multitasking while working on a study, such as watching TV, listening to music, and instant messaging. Zhou and Fishbach [68] reported that participant attrition was more pervasive yet less visible in online studies than in lab-based ones. The authors cautioned that failure to account for attrition rates, especially condition-dependent or selective dropout rates, would compromise the validity of findings from online studies. With regards to participants' domain knowledge, Tahaei and Vaniea [55] assessed the knowledge of online participants who self-claimed to have programming skills using five basic programming questions and compared their performance to computer science (CS) students. They found that while Prolific participants and CS students passed attention check questions at comparable rates, CS students were 26 times more likely to solve all programming questions than Prolific participants. Taken together, these findings indicate that further research is needed to better understand how online participants work on study tasks and how their performance is similar to or differs from participants recruited through traditional channels.

2.2 Reconceptualizing data quality in educational contexts

Central to our experiments is the notion of what it means to assess and compare the quality of data generated by research subjects from different recruitment channels. Previous descriptive research studies have typically relied on attention check questions embedded in online surveys (e.g., select "somewhat agree" for this question) to filter out inattentive respondents and low-quality data [31]. However, experienced online survey takers may be familiar with such attention check mechanisms and selectively respond to these questions with heightened attention. Additionally, recent research has discovered that attention check questions can be automatically answered through the malicious use of machine learning techniques [43].

To address the limitations of the attention check questions, this study employs more subtle and robust measurements to evaluate participants' engagement and performance through their interaction log data generated in a technology-based learning environment. We operationalize data quality as the comparability between online participants and a conventional sample of the target population, which in our case is college students in introductory STEM courses. The degree of comparability will be assessed through how participants engage with a cognitively demanding problem-solving task based on the interaction log data and whether they can obtain the correct solution.

In education research, student engagement is a multifaceted construct that has been the focus of a large body of literature [23, 27, 30, 53, 59]. Defined as the extent to which students are actively involved in learning activities [22], the construct is commonly conceptualized through three interrelated dimensions: behavioral, cognitive, and emotional [16]. For example, Stipek et al. [54] characterizes one type of active engagement as students taking on difficult tasks, exerting intense effort using deliberate problem-solving strategies, and persisting despite difficulty.

Previous research has shown that the behavioral aspect of engagement can be validly and efficiently measured using students' click-stream log data [30]. In particular, time-on-task derived from log data has been used as a proxy measure for behavioral engagement [3, 19, 24, 25, 33], and is often positively associated with learning outcomes [34, 56]. In problem-solving research, abnormally short time-on-task has been frequently linked with disengagement or deviating from the intended problem-solving behavior [1, 7, 8, 38, 66]. Given its empirical relevance to gauging engagement, we incorporated time-on-task as one measure for evaluating how well participants engage with the problem-solving tasks in our study. Furthermore, our previous research has validated the use of specific features extracted from interaction log data as indicators of problem-solving practices and strategies, thus offering a proxy measure for participants' cognitive engagement [61, 63, 65]. Taken together, these measures allow us to quantitatively compare the levels of engagement between the two groups of participants.

2.3 Review of problem-solving research

There is a long research tradition dedicated to the study of problem-solving in the physics education research (PER) community [5, 9, 15, 21, 29, 49]. Problem-solving situations arise as scientists investigate and build models and theories about the natural world and engineers design and build models and systems [32]. Authentic problems like these cannot be solved by recalling a formula based on pattern matching followed by fast and error-free calculation. Such problem-solving practice, also referred to as the "plug-and-chug" strategy, enables students to excel at end-of-chapter exercises and standardized exams but falls short in preparing them for complex, real-world challenges.

Over the past two decades, there has been growing recognition that STEM education should teach and measure students the practices and skills useful for solving authentic problems [36]. To this

end, a series of problem-solving activities have been developed in PhET Interactive Simulations (https://phet.colorado.edu/). These problems are designed to mimic the characteristics of authentic problems in science and engineering domains. Specifically, they require problem solvers to make decisions on the types of data to collect, the appropriate methods for obtaining and recording the data, and the relevant domain knowledge to apply to make sense of the collected data and arrive at a solution.

One of these problems is the Mystery Gift problem in the PhET Balancing Act simulation. The problem asks students to figure out the weight of a mystery gift using bricks with known weights and a balance scale that pivots around the center (Figure 1). Students can place the gift and bricks at various marked locations on the scale in the *Setup* mode and observe how it would rotate or stay balanced in the *Test* mode. To solve this problem, students need to balance the scale using the gift and bricks and apply the torque formula to calculate the weight of the gift. The simulation does not allow multiple bricks to be stacked at the same location. Furthermore, the weight of the gift was deliberately chosen to be unsolvable using a single brick. These features made the problem less intuitive and more difficult.

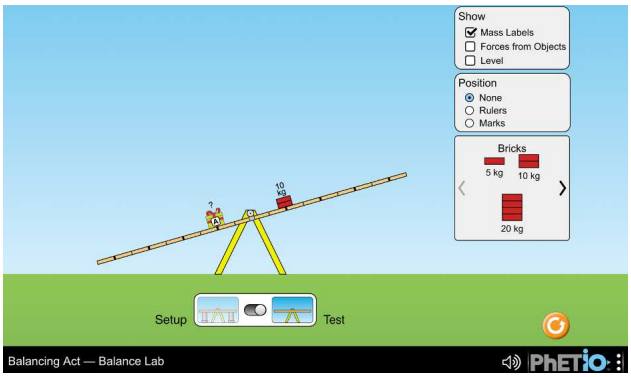

Figure 1: Screencast of the Mystery Gift problem. (Image by PhET Interactive Simulations, licensed under CC-BY 4.0)

Successfully solving the Mystery Gift problem hinges on the effective execution of the following problem-solving practices: data collection, data recording, and data interpretation. First, data collection involves gathering the data needed for calculating the weight of the mystery gift, including the weight of the brick(s) used and their respective distances from the central pivot point. This practice entails setting up experimental trials by placing the gift and bricks (5kg, 10kg, or 20kg) at various locations on the scale in the *Setup* mode and observing how the scale would rotate in the *Test* mode. Second, data recording involves maintaining a record of the collected data through note-taking. This practice can minimize the risk of losing track of the data from each test trial and facilitate subsequent problem-solving. Finally, data interpretation entails applying relevant domain knowledge, the torque formula in this case, to analyze and interpret the collected data and calculate the gift's weight. Notably, these practices have also been identified and discussed in previous research on scientific inquiry and problem-solving [40, 47, 50, 67].

Though there is no single correct solution path for solving the Mystery Gift problem, an expert-like approach entails systematically setting up simple trials for data collection and taking deliberate pauses from interacting with the simulation after each trial for data recording and interpretation. An effective problem solver would initially place a single brick and the mystery gift at equal distances on opposite sides of the pivot point to estimate the range of the gift's weight. In subsequent trials, a second brick would be deliberately added so that their combined torque gradually approaches the torque exerted by the gift, ultimately balancing the scale horizontally and making it possible to precisely calculate the gift's weight.

In terms of behavioral patterns extracted from the log data, effective data collection is evidenced by a problem solver consistently setting up simple test trials using no more than three objects. In contrast, setting up a large percentage of complex trials using four or more objects would signal a more random, trial-and-error approach to data collection. At the same time, pause after a trial, defined as the period when no interaction was logged in the simulation, serve as a critical indicator for evaluating the effectiveness of data recording and data interpretation.

We identified three types of pauses in our previous research analyzing the log data of students solving the Mystery Gift problem: mechanical, deliberate, and distracted [64]. Mechanical pauses are less than 10 seconds and generally account for the time it takes to view the animation of how the scale would rotate and move the mouse cursor with minimal cognitive processing involved. In contrast, deliberate pauses are between 10 seconds and three minutes and represent participants' deliberate efforts of stepping back from interacting with the simulation to work with the data collected and/or reflect on their progress. Lastly, distracted pauses are outliers in terms of duration (longer than three minutes) and signal off-task behaviors. This framework for categorizing pauses is aligned with and integrates findings from previous research examining student behavior in technology-based learning environments. For example, Paquette et al. [39] linked exceptionally short pauses between submission attempts to guessing behaviors in an intelligent tutoring system. Perez et al. [45] found that taking sufficiently long pauses when interacting with a circuit simulation was associated with better learning outcomes in an inquiry task. Gobert et al. [17] identified that rapidly running an educational simulation without pausing to think signaled disengagement from the task goals.

In summary, features extracted from the log data, such as the test trials set up and pauses after a trial, can serve as useful metrics for assessing specific problem-solving practices. These features allow for a more in-depth comparison of data quality across diverse groups of participants, encompassing both whether they can solve the problem (the product) and how they work through the problem (the process).

3 METHODS

3.1 Materials and procedures

Two instances of the mystery gift problem were used for this study. These two problems exhibit comparable levels of difficulty as confirmed by empirical pilot testing. Participants were first given a real-world context of the problem, that they needed to weigh a gift in order to print out a shipping label to mail the gift at home. Subsequently, participants viewed a brief tutorial on how to use the PhET Balancing Act simulation. They were then instructed to solve the weight of the first gift within 15 trials as a practice round. Afterward, participants were provided with a worked-out solution to the practice problem. Finally, participants engaged in the test round, in which they were asked to solve the weight of a different gift within 15 trials. The study has been reviewed and approved by the Institutional Review Boards (IRBs) at the authors' institutions (Protocol ID: 29325).

3.2 Participants

3.2.1 University. In the first experiment, we recruited students from three parallel sections of a calculus-based college introductory level physics course in a large public university in the US. Of the students enrolled in the course, 24% were female, 35% were from racial or ethnic groups traditionally underrepresented in STEM disciplines, 13% were first-generation students, and the average age was 20. The majority of students enrolled in this class were from engineering or computer science majors. The mystery gift problem was presented as an extra credit activity to students, using the Obojobo Learning Objects Platform [14] and connected via Learning Tools Interoperability (LTI) to the Canvas Learning Management System (LMS). Students completed the study at a time and location of their choice and received a small amount of extra credit for their participation. Students would receive the full extra credit if they correctly solved the test problem and 90% of the extra credit if their answer was incorrect. Students enrolled in the course have studied the concept of torque and torque balance a few weeks prior to accessing the task. The final University sample includes 325 students (29% female).

3.2.2 Prolific. In the second experiment, we recruited 40 participants from Prolific and used the platform's built-in pre-screener to reach the target population. Our inclusion criteria were current undergraduate students located in the US, majoring in a STEM subject, and having not taken part in previous Prolific studies run by our research team. We also requested the sample to be gender balanced. The majority of the participants (70%) fell within the 18-21 age range, followed by those aged 22-25 (15%), 26-29 (10%), and 30 and above (5%). 92.5% of the participants came from a four-year university/college, and 7.5% came from a community college.

The two mystery gift problems, along with the other questions, were embedded in an online Qualtrics survey. Participants completed the study at a time and location of their choice and were compensated $4 for their participation in the 30-min study. They also had the opportunity to get a $2 bonus for correctly solving the mystery gift in the test round. Prior to introducing participants to the problem-solving tasks, we posed two questions to gauge their background knowledge in physics. The first question asks participants to self-rate how familiar they are with torque, or a force that causes rotations, choosing from three levels: not familiar at all/having a conceptual understanding/knowing the exact formula. The second question asks participants to identify the correct torque formula from four options. Based on an analysis of the log data, four participants completed the survey but did not use the mystery gift when interacting with the simulation and were subsequently excluded from data analyses. The final Prolific sample includes 36 participants (47% female).

3.3 Data processing and analyses

To evaluate participants' engagement levels, we calculated their time spent on solving the problems based on the first and last timestamps in the log data and the time spent on viewing the solution to the practice problem as recorded by the Qualtrics survey or LMS platform. To measure participants' outcomes on the problem-solving task, we scored their answers to the gift's weight into two levels: correct and incorrect. Participants in the correct group submitted answers that came within one kilograms of the gift's actual weight (e.g., an answer between 54.01kg and 55.99kg would be correct when the gift weighs 55kg), while participants in the incorrect group submitted answers that were outside of the acceptable range. Differences in solution quality between the Prolific and University samples were tested using the chi-square test of independence.

To quantitatively characterize participants' work processes and problem-solving practices/strategies, we wrote a Python script to process the Javascript Object Notation (JSON) files recording participants' interactions. Three behavioral features were extracted from individual participants' log data: 1) total number of trials set up when solving the problem: 2) frequency of different types of trials (simple vs. complex), and 3) frequency of different types of pauses (mechanical, deliberate, or distracted) after setting up each trial.

While time-on-task provides a general account of participants' levels of engagement, the trial- and pause-based features can help us develop a more nuanced understanding of how individual participants approached the task. Distributions of continuous variables were assessed using the Shapiro-Wilk test for normality. Two-sample t-tests or Mann-Whitney U tests were then used to compare each feature between the Prolific and University participants. A Bonferroni adjusted alpha level was adopted to correct for multiple comparisons.

4 RESULTS

We progress through the results in the following order to elucidate the comparisons between the two groups of participants: 1) dropout rates (i.e., participants who signed up but did not finish the study); 2) physics knowledge background; 3) engagement levels as measured by their time spent on the tasks; 4) solution qualities; 5) problem-solving processes as captured by various behavioral features extracted from the log data.

4.1 Dropout rates

The dropout rate was calculated as the percentage of participants who accessed yet did not complete the study. Among the 54 Prolific participants who accessed the study, 14 (26%) terminated their participation without completion. For the University study, a total of 500 students opened the online module containing the study. Among those students, 325 worked on both the practice and test problems as recorded by the log data with matching IDs and submitted their answers, resulting in a dropout rate of 35%. The difference in dropout rates between the two samples was not statistically significant (chi-squared (1) = 0.23, $p = 0.63$).

4.2 Knowledge Background

There was a large variation in Prolific participants' background knowledge in physics. 31% (11 out of 36) of the participants self-reported that they were not at all familiar with the concept of torque at the beginning of the online survey. 58% reported having a conceptual understanding and 11% reported that they knew the exact torque formula. In a separate multiple-choice question, 39% of the Prolific participants correctly identified the torque formula from a list of four options.

In contrast, all the University students were recruited from a calculus-based introductory physics course. Torque and rotational kinematics were covered in the course over a period of two weeks before the study was made available to students. Therefore, it is reasonable to expect that all University participants have at least been exposed to the concept of torque and the torque equation. To further evaluate University students' knowledge in torque, we analyzed their performance on two questions that explicitly addressed torque in the midterm exam in one section of the course. 78% of the students managed to correctly solve at least one of the torque questions.

4.3 Time-on-task

To understand how participants engaged with the problem-solving tasks in the study, we compared their time spent on solving the problems and viewing the solution to the practice problem. The Mann-Whitney U tests were used due to non-normal distribution of time-on-task. We found no significant difference in the time spent on solving the practice problem, but Prolific participants overall worked longer on the test problem than University participants. The median time spent on the practice problem was 3.15 mins (IQR: 1.60 – 6.37 mins) for the University sample and 3.69 mins (IQR: 2.61 – 6.76 mins) for the Prolific sample. This difference was not statistically significant ($p = 0.25$). On the other hand, University students spent a median time of 1.33 mins (IQR: 0.65 – 3.37 mins) while Prolific participants spent a median time of 3.31 mins (IQR: 1.10 – 5.00 mins) on the test problem ($p = 0.003$).

To better understand how participants spent their time solving the test problem, we divided participants in each sample into two levels based on their solution quality. Figure 2 presents the boxplots of individual participants' time spent on solving the test problem grouped by sample and solution quality. We did not find a group effect (Prolific vs. University) in the time spent on solving the test problem among the participants who obtained the correct solution. The significant difference in time-on-task was largely driven by participants who did not solve the test problem. While Prolific participants who did not reach a correct solution spent a median of 4.27 mins (IQR: 1.15 – 5.32 mins) on the test problem, the median time-on-task was only 1.07 mins (IQR: 0.59 – 2.15 mins) for University students at the same solution level (Mann-Whitney U test, $p < 0.001$). This result suggests that there was a divergent pattern of engagement among participants who did not solve the test problem, with Prolific participants demonstrating a higher level of engagement with the task than University participants.

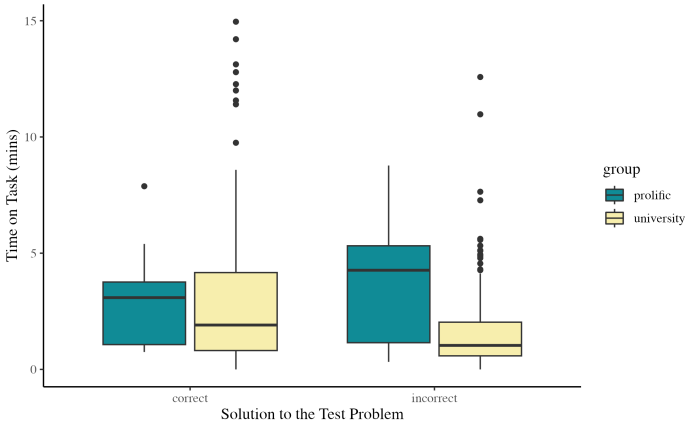

Figure 2: Time spent on solving the test problem grouped by sample and solution quality

4.4 Time spent on viewing the solution to the practice problem

All participants were given a worked example describing how to solve the practice problem step-by-step after attempting to solve the problem on their own. There was no significant difference between the two groups of participants in terms of the time spent on viewing the solution. The median viewing time was 1.75 mins (IQR: 1.17 – 2.75 mins) for the University sample and 1.54 mins (IQR: 0.96 – 2.21 mins) for Prolific (Mann-Whitney U test, $p = 0.12$). We also did not find any difference in the viewing time between University and Prolific participants who solved the test problem or between those who did not solve the test problem.

4.5 Solution quality

To determine the comparability of problem-solving outcomes between University and Prolific participants, we compared the percentage of correct solutions across the two groups. Table 1 presents the percentage and count of participants who correctly solved the gift's weight in the practice and test problems. 38% of the University participants and 28% of the Prolific participants solved the practice problem. A chi-square test of independence showed that there was no significant association between group membership (University vs. Prolific) and obtaining the correct solution (chi-square (1) = 1.01, $p = 0.31$). For the test problem, the same percentage (42%) of participants in the University and Prolific samples obtained the correct solution. This result indicates that despite differences in

their physics background knowledge, University and Prolific participants performed about the same in terms of obtaining the correct answers for the mystery gift problems.

Table 1: Percentage and count of correct solutions

	University	Prolific
Mystery Gift I (Practice)	38% (123)	28% (10)
Mystery Gift II (Test)	42% (136)	42% (15)

4.6 Number of trials

Overall, we did not find any significant difference in the number of trials set up in the practice or test problem across the University and Prolific samples. The median number of trials was 14 for both the University and Prolific participants in the practice task. This is not surprising as the instruction asked students to solve the problem in 15 trials or less.

At the same time, we found that a substantial number of participants continued working on the problem despite reaching the trial limit, as the simulation does not have a built-in mechanism to stop the task once a certain trial count has been reached. 41% of University students and 39% of Prolific participants set up more than 15 trials when solving the practice problem. One exceptional University student set up 298 trials and finally reached a correct solution.

For the test problem, the median number of trials was 8 for University participants and 10 for Prolific participants, though this difference was not statistically significant (Mann-Whitney U test, $p = 0.06$). The significant difference in time-on-task for the test problem, as presented in an earlier section of the results, yet lack of corresponding difference in the number of trials set up indicates that the University and Prolific participants worked on the test problem at a different pace, a key result to be discussed in a subsequent section.

4.7 Percentage of complex trials

The percentage of complex trials is an indicator of how effective a problem solver is at collecting data, with higher percentages of complex trials (i.e., using four or more objects in a trial) indicating less effective data collection practice. We found no significant difference in the percentage of complex trials set up when solving the practice (Mann-Whitney U test, $p = 0.90$) or test problem (Mann-Whitney U test, $p = 0.66$) across the University and Prolific samples. For participants in both groups, the median percentage of complex trials was around 25% in the practice problem and 40% in the test problem. The high percentage of complex trials set up indicates that the trial-and-error approach of adding an increasing number of objects in the hope of balancing the scale was a popular problem-solving strategy for participants in both groups.

4.8 Percentage of deliberate pauses

The percentage of deliberate pauses ($10\ secs \leq duration \leq 3\ mins$) during individual participant's problem-solving process indicates the pace at which participants worked, with a higher percentage corresponding to a more deliberate approach for data recording and data interpretation. Results of the Mann-Whitney U tests showed that Prolific participants had a higher percentage of deliberate pauses than University students when solving the test problem ($p < 0.001$). Furthermore, we found a similar pattern as the result of the time-on-task analysis, that the difference between the two groups was primarily driven by low-performing students who did not reach a correct solution for the test problem.

While there was no significant difference in the percentage of deliberate pauses between the high-performing students across the two samples in either the practice or test problem, University participants who did not obtain the correct solution for the test problem had a significantly lower percentage of deliberate pauses than Prolific participants at the same solution level (Figure 3). The median percentage of deliberate pauses was 0% for low-performing University participants, indicating that a substantial proportion of those participants did not make any deliberate pause after setting up a test trial. In contrast, the median percentage of deliberate pauses when solving the test problem was 27% for low-performing Prolific participants. The significant difference in the percentage of deliberate pauses suggests that low-performing University students made only cursory attempts to solve the test problem compared to Prolific participants.

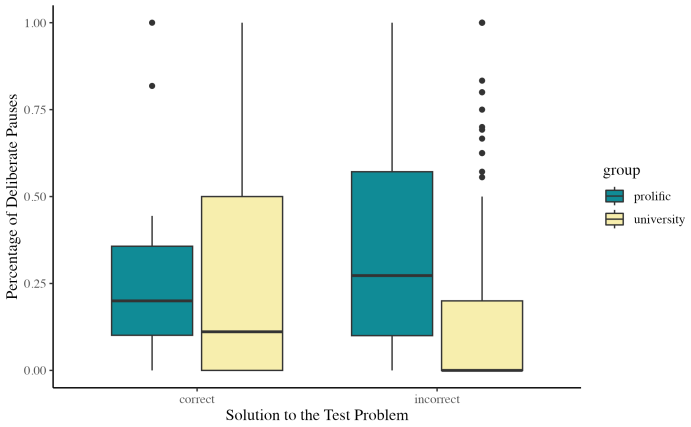

Figure 3: Percentage of deliberate pauses grouped by sample and solution quality

We also examined a different type of pause as captured by the log data: distracted pauses or pauses that are longer than three minutes. We found that the occurrence of such pauses was overall low in both the University and Prolific samples, with a total of 39 instances of distracted pauses belonging to 30 students (9%) in the University sample and three instances belonging to three participants in the Prolific sample (8%). This result implies that neither group of participants frequently engaged in off-task behaviors during the study.

To summarize, our study finds that Prolific participants who passed the pre-screener as undergraduate STEM major students performed at a similar level in terms of solution quality on an interactive physics problem as students in an introductory physics course. Table 2 summarizes the results of key comparisons. We

found significant differences between the University and Prolific participants in only two log data-based behavioral features: the time on task and the percentage of deliberate pauses when solving the test problem.

A revealing outcome of this study is the behavioral difference between the low-performing participants who submitted incorrect solutions for the test problem. While the University students who did not obtain the correct solution spent a short time on the test problem with minimal deliberate pauses, Prolific participants were more deliberate and worked longer on the problem. In contrast, there was no significant differences in the engagement level or problem-solving process between the high-performing participants who reached the correct solutions in University and Prolific samples.

5 DISCUSSION

5.1 Summary of results

This study examined the general comparability of data obtained through the Prolific and University subject pools using an interactive problem-solving task embedded in a physics simulation. In contrast to questionnaires and surveys aimed at capturing people's beliefs and perceptions, the tasks used in the current study are cognitively demanding and require the application of physics knowledge as well as effective problem-solving practices. Furthermore, rather than relying on attention check questions at arbitrary time points during a study, we extracted a detailed picture of individual participants' problem-solving process based on the log files of their interactions in the task environment. The log data provides key indicators of the problem-solving practices adopted by participants and makes it possible to detect and remove invalid responses (e.g., the four participants who did not use the mystery gift when working in the simulation yet submitted answers for the gift's weight in the survey).

For RQ1, we found that Prolific participants self-reported more varied and generally less robust levels of physics background knowledge compared to the University students enrolled in a physics course. At the same time, we did not find any significant difference in the level of behavioral engagement as measured by time-on-task between high-performing Prolific and University participants who reached the correct answer. On the other hand, University participants who did not solve the problem exhibited a lower level of behavioral engagement compared to their Prolific counterparts.

For RQ2, we found no evidence of significant differences in the problem-solving outcomes as measured by solution quality between the two groups of participants, despite the Prolific participants self-reporting more varied and generally less physics background knowledge. For RQ3, behavioral features extracted from log data indicated that both groups of participants adopted comparable problem-solving practices for data collection, as evidenced by the number and types of test trials set up. However, a notable divergence was observed for low-performing University students who did not reach the correct solution. This subgroup was more likely to rush through the task without any deliberate efforts to record or interpret the data collected, as they made significantly fewer deliberate pauses during the problem-solving process than Prolific participants at the same solution quality level.

5.2 Connections with prior research

Findings from the current study are in line with the results of previous studies that the data obtained from online research platforms were comparable to those obtained from traditional subject pools [20, 28, 41, 60]. Moreover, we did not find evidence for the claim that the responses of online participants were low-quality due to their inattentiveness or being prone to distraction [10, 11].

At the same time, this study demonstrates the utility of log data generated in interactive learning environments for unobtrusively capturing the problem-solving processes across diverse groups of individuals [62]. The finding that an absence of sufficiently long pauses during problem-solving is associated with poor performance is in line with previous research on student behavior in technology-based learning environments [2, 39, 45]. In addition, the study also opens up new research directions for how we teach problem-solving: behavioral features extracted from the log data can guide educators in diagnosing failure modes and designing personalized instructional interventions to enhance students' problem-solving competency.

Furthermore, our findings build on and extend previous research on how time-on-task can serve as a metric for assessing student engagement [3, 8, 34]. Specifically, we observed that the subgroup of University students who failed to solve the Mystery Gift problem also had the shortest time-on-task. However, our findings also caution that high levels of behavioral engagement alone do not guarantee success in problem-solving tasks, as the subgroup of Prolific participants who were engaged based on time-on-task still struggled to solve the problem. These results underscore the complex interactions among engagement, domain knowledge, and problem-solving strategies and practices that affect students' success in solving authentic problems.

5.3 Interpretation of results

How did the Prolific sample achieve problem-solving outcomes comparable to those of the University sample despite reporting less formalized knowledge of torque? One explanation lies in the real-world and interactive characteristics of the Mystery Gift problem. The problem differs from textbook-style questions that students regularly practice in physics courses. Instead, it is more similar to real-world problems and requires students to adopt a series of effective problem-solving and decision-making practices to solve the problem, including collecting data through interacting with the simulation and applying relevant knowledge at the right time to interpret the data collected [46, 50]. Furthermore, the interactive nature of the simulation allows participants to build on their intuitive understanding of factors affecting rotational force gained through everyday experiences like playing on a seesaw and bootstrap a formal understanding of torque, which in turn provides the domain knowledge needed for solving the problem at hand.

The difference in time-on-task also imply that low-performing Prolific participants devoted more effort to the problem-solving task than their University peers. Why did low-performing Prolific and University participants exhibit divergent engagement patterns when working on the same problem-solving task? We postulate that the underlying motivations and contexts substantially influenced their respective approaches to the problem.

Table 2: Key characteristics and comparisons of the university and Prolific samples

	University	Prolific
Dropout Rate	35%	26%
Mystery Gift I (Practice)		
Time-on-task	3.15 mins [1.60 – 6.37]	3.69 mins [2.61 – 6.76]
Number of Trials	14 [9 - 23]	14 [9 - 21]
Percentage of Complex Trials	25% [0 – 50%]	28% [0 – 43%]
Percentage of Deliberate Pauses	13% [6% - 29%]	29% [11% - 53%]
Time on viewing the solution	1.75 mins [1.17 – 2.75]	1.54 mins [0.96 – 2.21]
Mystery Gift II (Test)		
Time-on-task*	1.33 mins [0.65 – 3.37]	3.31 mins [1.10 – 5.00]
Number of Trials	8 [6 - 12]	10 [8 - 12]
Percentage of Complex Trials	44% [11% – 64%]	41% [0 – 67%]
Percentage of Deliberate Pauses*	0% [0 – 33%]	24% [10% – 47%]

* indicates significant differences after Bonferroni adjustment

First, the reward structure differed for the two groups, potentially altering their performance dynamics. University students received course credits, which is a common method employed in academic settings to recruit students for research studies. This may unintentionally contributed to a satisficing behavior where minimal effort was sufficient to gain most of the rewards [26, 52]. In the context of solving the Mystery Gift problem, students may not have sufficient motivation to work hard to obtain the correct solution when simply attempting to solve the problem and submitting a wrong answer would give them 90% of the allotted course credits. Prolific participants, on the other hand, received a small amount of monetary rewards tied to performance, which has been shown to improve effort and outcomes in tasks like Bayesian reasoning compared to flat-fee incentives and course credits [4].

Second, Prolific's approval rating system created an additional layer of motivation for its participants. Participants might risk having their submissions rejected if they complete them too hastily, according to Prolific's criteria [48]. Each Prolific participant has an approval rating calculated as the number of approved submissions divided by the total number of submissions. As some studies would only recruit participants with an approval rating of 95% or higher, participants are motivated to put in the effort and avoid their submissions being rejected. This is unlike the University context, where students face no such approval rating or reputational risk, possibly affecting their level of engagement in the study.

Finally, the timing of the study offers another contextual layer for understanding these differing behaviors. While the Prolific study did not coincide with any particular period, the University study was conducted during the final exam study period, a time when students typically have a heavy workload with final assignments and exams. Previous research indicates a notable drop in intrinsic motivation and performance in such academic conditions compared to early in the semester, specifically for those receiving course credits [35]. It is possible that a subset of University students experienced low levels of intrinsic motivation or limited availability due to the timing of the study, preventing them from fully engaging in the problem-solving tasks.

5.4 Limitations and Future Research

One limitation of the current study is the differing proportions of female students between the two samples, which may influence the results. On the other hand, the ability to recruit a higher percentage of female participants through crowdsourcing platforms may be a notable advantage, as this can help educational researcher better understand female students' experiences in STEM education. Another limitation not fully addressed by this study is the sampling bias due to online participants "self-selecting" into the study or dropping out of the study without completion [31, 68]. We did not collect data on the reasons for choosing to enroll in this specific study from Prolific participants. We also have no insights into why a total of 14 (26%) participants dropped out of the study as the platform precludes researchers from accessing the information of participants who exited the study without completion. It is possible that only students who are interested in physics and/or have a high self-efficacy in physics chose to participate and finish the study in the first place. We should note that this potential sampling bias also applies to traditional studies, as we observed a dropout rate of 35% from the University participants. Future studies should investigate what motivates students to participate in studies that employ cognitively demanding tasks and what the considerations are for choosing to drop out without completion.

6 CONCLUSION

This study lends support to the view that online crowdsourcing research platforms "provide education researchers with a workable complement to traditional sampling methods and may be particularly applicable for research whose aim is to study characteristics of college-aged and adult learners" [13]. There are several advantages to conducting education research on crowdsourcing platforms, including a streamlined and simplified data collection cycle, the ability to efficiently obtain a sample from the target population using prescreeners, and the capacity to support various experimental designs such as randomization and longitudinal studies by integrating with online survey platforms. Our findings also indicate that there are several important practices that should be taken

into account to ensure data quality when conducting educational research on Prolific. These practices include 1) incorporating questions to gauge participants' background knowledge, thus allowing for a more precise contextualization of the study's findings; and 2) deploying technology-based learning environments and learning analytics techniques to capture nuanced behavioral indicators for participants' work processes, such as time-on-task and other log data-based features. With careful study design and data-rich digital task environments, crowdsourcing platforms can be useful for empirical research in education, allowing for piloting innovative instructional and assessment tasks as well as examining the efficacy of specific educational interventions efficiently and at scale.

ACKNOWLEDGMENTS

This study was partially funded by NSF Grant No. DUE-1845436.

REFERENCES

[1] Giora Alexandron, José A Ruipérez-Valiente, Zhongzhou Chen, Pedro J Muñoz-Merino, and David E Pritchard. 2017. Copying@ Scale: Using harvesting accounts for collecting correct answers in a MOOC. *Computers & Education* 108 (2017), 96–114.

[2] Saleema Amershi, Cristina Conati, et al. 2009. Combining unsupervised and supervised classification to build user models for exploratory learning environments. *Journal of educational data mining* 1, 1 (2009), 18–71.

[3] Roger Azevedo. 2015. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational psychologist* 50, 1 (2015), 84–94.

[4] Gary L Brase. 2009. How different types of participant payments alter task performance. *Judgment and Decision making* 4, 5 (2009), 419–428.

[5] EW Burkholder, JK Miles, TJ Layden, KD Wang, AV Fritz, and CE Wieman. 2020. Template for teaching and assessment of problem solving in introductory physics. *Physical Review Physics Education Research* 16, 1 (2020), 010123.

[6] Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46 (2014), 112–130.

[7] Zhongzhou Chen. 2022. Measuring the level of homework answer copying during COVID-19 induced remote instruction. *Physical Review Physics Education Research* 18, 1 (2022), 010126.

[8] Zhongzhou Chen, Mengyu Xu, Geoffrey Garrido, and Matthew W Guthrie. 2020. Relationship between students' online learning behavior and course performance: What contextual information matters? *Physical Review Physics Education Research* 16, 1 (2020), 010138.

[9] Michelene TH Chi, Paul J Feltovich, and Robert Glaser. 1981. Categorization and representation of physics problems by experts and novices. *Cognitive science* 5, 2 (1981), 121–152.

[10] Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473.

[11] Scott Clifford and Jennifer Jerit. 2014. Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science* 1, 2 (2014), 120–131.

[12] Margaret Anne Defeyter, Paul B Stretesky, Michael A Long, Sinéad Furey, Christian Reynolds, Debbie Porteous, Alyson Dodd, Emily Mann, Anna Kemp, James Fox, et al. 2021. Mental well-being in UK higher education during COVID-19: Do students trust universities and the government? *Frontiers in public health* 9 (2021), 646916.

[13] D Jake Follmer, Rayne A Sperling, and Hoi K Suen. 2017. The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher* 46, 6 (2017), 329–334.

[14] Center for Distributed Learning. 2023. Obojobo. https://next.obojobo.ucf.edu/ Available from: https://next.obojobo.ucf.edu/.

[15] David Fortus. 2009. The importance of learning to make assumptions. *Science Education* 93, 1 (2009), 86–108.

[16] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[17] Janice D Gobert, Ryan S Baker, and Michael B Wixon. 2015. Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist* 50, 1 (2015), 43–57.

[18] Joseph K Goodman and Scott Wright. 2022. MTurk and online panel research: The impact of COVID-19, bots, TikTok, and other contemporary developments. (2022).

[19] 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*. 41–50.

[20] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48 (2016), 400–407.

[21] Joan I Heller and Frederick Reif. 1984. Prescribing effective human problem-solving processes: Problem description in physics. *Cognition and instruction* 1, 2 (1984), 177–216.

[22] Sue Helme and David Clarke. 2001. Identifying cognitive engagement in the mathematics classroom. *Mathematics Education Research Journal* 13, 2 (2001), 133–153.

[23] Min Hu and Hao Li. 2017. Student engagement in online learning: A review. In *2017 International Symposium on Educational Technology (ISET)*. IEEE, 39–43.

[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*. 31–40.

[25] Vitomir Kovanović, Dragan Gašević, Shane Dawson, Srećko Joksimović, Ryan S Baker, and Marek Hatala. 2015. Penetrating the black box of time-on-task estimation. In *Proceedings of the fifth international conference on learning analytics and knowledge*. 184–193.

[26] Jon A Krosnick, Sowmya Narayan, and Wendy R Smith. 1996. Satisficing in surveys: Initial evidence. *New directions for evaluation* 1996, 70 (1996), 29–44.

[27] George D Kuh. 2009. The national survey of student engagement: Conceptual and empirical foundations. *New directions for institutional research* 141 (2009), 5–20.

[28] Alexia Micallef and Philip M Newton. 2022. The Use of Concrete Examples Enhances the Learning of Abstract Concepts; A Replication Study. *Teaching of Psychology* (2022), 00986283211058069.

[29] Jeff Milbourne and Eric Wiebe. 2018. The role of content knowledge in ill-structured problem solving for high school physics students. *Research in Science Education* 48 (2018), 165–179.

[30] Benjamin Motz, Joshua Quick, Noah Schroeder, Jordon Zook, and Matthew Gunkel. 2019. The validity and utility of activity logs as a measure of student engagement. In *Proceedings of the 9th international conference on learning analytics & knowledge*. 300–309.

[31] Alexander Newman, Yuen Lam Bavik, Matthew Mount, and Bo Shao. 2021. Data collection via online platforms: Challenges and recommendations for future research. *Applied Psychology* 70, 3 (2021), 1380–1402.

[32] NGSS. 2013. Next generation science standards: For states, by states. (2013).

[33] Quan Nguyen. 2020. Rethinking time-on-task estimation with outlier detection accounting for individual, time, and task differences. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 376–381.

[34] Quan Nguyen, Michal Huptych, and Bart Rienties. 2018. Linking students' timing of engagement to learning design and academic performance. In *Proceedings of the 8th international conference on learning analytics and knowledge*. 141–150.

[35] Michael ER Nicholls, Kellie M Loveless, Nicole A Thomas, Tobias Loetscher, and Owen Churches. 2015. Some participants may be better than others: Sustained attention and motivation are higher early in semester. *Quarterly journal of experimental psychology* 68, 1 (2015), 10–18.

[36] NRC. 2012. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

[37] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[38] David J Palazzo, Young-Jin Lee, Rasil Warnakulasooriya, and David E Pritchard. 2010. Patterns, correlates, and reduction of homework copying. *Physical Review Special Topics-Physics Education Research* 6, 1 (2010), 010104.

[39] Luc Paquette, Adriana MJB de Carvalho, and Ryan Shaun Baker. 2014. Towards Understanding Expert Coding of Student Disengagement in Online Learning.. In *CogSci*.

[40] Margus Pedaste, Mario Mäeots, Leo A Siiman, Ton De Jong, Siswa AN Van Riesen, Ellen T Kamp, Constantinos C Manoli, Zacharias C Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review* 14 (2015), 47–61.

[41] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology* 70 (2017), 153–163.

[42] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2022), 1.

[43] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue. 2020. Attention please: Your attention check questions in survey studies can be automatically answered. In *Proceedings of The Web Conference 2020*. 1182–1193.

[44] Deanna Perez, Janelle K Thalken, Nzubechukwu E Ughelu, Camilla J Knight, and William V Massey. 2021. Nowhere to go: Parents' descriptions of children's physical activity during a global pandemic. *Frontiers in Public Health* 9 (2021),

642932.
[45] Sarah Perez, Jonathan Massey-Allard, Deborah Butler, Joss Ives, Doug Bonn, Nikki Yee, and Ido Roll. 2017. Identifying productive inquiry in virtual labs using sequence mining. In *Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18*. Springer, 287–298.
[46] Argenta Price, Shima Salehi, Eric Burkholder, Candice Kim, Virginia Isava, Michael Flynn, and Carl Wieman. 2022. An accurate and practical method for assessing science and engineering problem-solving expertise. *International Journal of Science Education* 44, 13 (2022), 2061–2084.
[47] Argenta M Price, Candice J Kim, Eric W Burkholder, Amy V Fritz, and Carl E Wieman. 2021. A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment. *CBE—Life Sciences Education* 20, 3 (2021), ar43.
[48] Prolific. 2023. Approvals, rejections, returns. https://researcher-help.prolific.co/hc/en-gb/articles/360009092394-Approvals-rejections-returns Accessed: 2023-09-01.
[49] Frederick Reif and Joan I Heller. 1982. Knowledge structure and problem solving in physics. *Educational psychologist* 17, 2 (1982), 102–127.
[50] Shima Salehi. 2018. *Improving problem-solving through reflection*. Stanford University.
[51] Laura Scheel, Gergana Vladova, and André Ullrich. 2022. The influence of digital competences, self-organization, and independent learning abilities on students' acceptance of digital learning. *International journal of educational technology in higher education* 19, 1 (2022), 1–33.
[52] Herbert A Simon. 1957. Models of man; social and rational. (1957).
[53] Gale M Sinatra, Benjamin C Heddy, and Doug Lombardi. 2015. The challenges of defining and measuring student engagement in science. , 13 pages.
[54] Deborah J Stipek et al. 1996. Motivation and instruction. *Handbook of educational psychology* 1 (1996), 85–113.
[55] Mohammad Tahaei and Kami Vaniea. 2022. Recruiting participants with programming skills: A comparison of four crowdsourcing platforms and a CS student mailing list. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
[56] Dirk T Tempelaar, Bart Rienties, and Bas Giesbers. 2015. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior* 47 (2015), 157–167.
[57] Anne M Turner, Thomas Engelsma, Jean O Taylor, Rashmi K Sharma, and George Demiris. 2020. Recruiting older adult participants through crowdsourcing platforms: Mechanical Turk versus Prolific Academic. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 1230.
[58] Nina Vaupotič, Dorothe Kienhues, and Regina Jucks. 2022. Gaining insight through explaining? How generating explanations affects individuals' perceptions of their own and of experts' knowledge. *International Journal of Science Education, Part B* 12, 1 (2022), 42–59.
[59] Jovita M Vytasek, Alexandra Patzak, and Philip H Winne. 2020. Analytics for student engagement. *Machine learning paradigms: Advances in learning analytics* (2020), 23–48.
[60] Sheryl L Walter, Scott E Seibert, Daniel Goering, and Ernest H O'Boyle. 2019. A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology* 34 (2019), 425–452.
[61] Karen Wang, Krishnan Nair, and Carl Wieman. 2021. Examining the links between log data and reflective problem-solving practices in an interactive task. In *LAK21: 11th international learning analytics and knowledge conference*. 525–532.
[62] Karen D Wang, Jade Maï Cock, Tanja Käser, and Engin Bumbacher. 2023. A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. *British Journal of Educational Technology* 54, 1 (2023), 192–221.
[63] Karen D Wang, Shima Salehi, Max Arseneault, Krishnan Nair, and Carl Wieman. 2021. Automating the assessment of problem-solving practices using log data and data mining techniques. In *Proceedings of the eighth ACM conference on learning@ scale*. 69–76.
[64] Karen D Wang, Shima Salehi, and Carl Wieman. 2023. Applying Log Data Analytics to Measure Problem Solving in Simulation-Based Learning Environments. In *Unobtrusive Observations of Learning in Digital Environments: Examining Behavior, Cognition, Emotion, Metacognition and Social Processes Using Learning Analytics*. Springer, 31–52.
[65] Karen D Wang and Carl Wieman. 2022. Applying Sequence Mining to Explore Students' Problem-Solving Practices Using an Interactive Simulation-Based Task. In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 433-439*. International Society of the Learning Sciences.
[66] Rasil Warnakulasooriya, David J Palazzo, and David E Pritchard. 2007. Time to completion of web-based physics problems with tutoring. *Journal of the experimental analysis of behavior* 88, 1 (2007), 103–113.
[67] Mark Windschitl, Jessica Thompson, and Melissa Braaten. 2008. Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science education* 92, 5 (2008), 941–967.
[68] Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology* 111, 4 (2016), 493.