Am I Wrong, or Is the Autograder Wrong? Effects of AI Grading Mistakes on Learning

Tiffany Wenting Li*
Silas Hsu*
wenting7@illinois.edu
silash2@illinois.edu
Department of Computer Science,
University of Illinois at
Urbana-Champaign
United States

Max Fowler
mfowler5@illinois.edu
Department of Computer Science,
University of Illinois at
Urbana-Champaign
United States

Zhilin Zhang
zhilin.zhang@cs.ox.ac.uk
Department of Computer Science,
University of Oxford
United Kingdom

Craig Zilles
zilles@illinois.edu
Department of Computer Science,
University of Illinois at
Urbana-Champaign
United States

Karrie Karahalios
kkarahal@illinois.edu
Department of Computer Science,
University of Illinois at
Urbana-Champaign
United States

ABSTRACT

Errors in AI grading and feedback often have an intractable set of causes and are, by their nature, difficult to completely avoid. Since inaccurate feedback potentially harms learning, there is a need for designs and workflows that mitigate these harms. To better understand the mechanisms by which erroneous AI feedback impacts students' learning, we conducted surveys and interviews that recorded students' interactions with a short-answer AI autograder for "Explain in Plain English" code reading problems. Using causal modeling, we inferred the learning impacts of wrong answers marked as right (false positives, FPs) and right answers marked as wrong (false negatives, FNs). We further explored explanations for the learning impacts, including errors influencing participants' engagement with feedback and assessments of their answers' correctness, and participants' prior performance in the class.

FPs harmed learning in large part due to participants' failures to detect the errors. This was due to participants not paying attention to the feedback after being marked as right, and an apparent bias against admitting one's answer was wrong once marked right. On the other hand, FNs harmed learning only for survey participants, suggesting that interviewees' greater behavioral and cognitive engagement protected them from learning harms. Based on these findings, we propose ways to help learners detect FPs and encourage deeper reflection on FNs to mitigate the learning harms of AI errors.

*Both authors contributed equally to this research.

ICER '23 V1, August 07–11, 2023, Chicago, IL, USA

ACM ISBN 978-1-4503-9976-0/23/08…$15.00
https://doi.org/10.1145/3568813.3600124

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Interactive learning environments**; • **Social and professional topics** → **Computing education**.

KEYWORDS

human-AI interaction, AI error, formative feedback, autograder, computer science education, automated short answer grading, explain in plain English, EiPE, Bayesian modeling

ACM Reference Format:
Tiffany Wenting Li, Silas Hsu, Max Fowler, Zhilin Zhang, Craig Zilles, and Karrie Karahalios. 2023. Am I Wrong, or Is the Autograder Wrong? Effects of AI Grading Mistakes on Learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1 (ICER '23 V1), August 07–11, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3568813.3600124

1 INTRODUCTION

Formative feedback is a critical facilitator of learning and performance [4, 5, 20, 34, 60]. Timely feedback especially helps students learn [70], and computers are well-positioned to provide prompt feedback when large course sizes would make it otherwise impractical, such as in large university courses and Massive Open Online Courses (MOOCs). With advancements in Natural Language Processing (NLP), computers can now quickly provide grading and feedback on essays and short answer responses [28, 43]. This facilitates the scaling up of courses that have traditionally relied on these types of assessments, as well as novel assessment strategies in computing education, such as the "Explain in Plain English" (EiPE) code reading problems that form the context of this paper. However, such feedback, including AI-generated ones, is sometimes inaccurate.

Past studies have demonstrated that on simple tasks where computers can provide perfectly accurate feedback (e.g., multiple choice), people were less likely to master a task when researchers

intentionally gave them inaccurate feedback [12, 32, 36]. But no prior work has addressed the learning impacts of erroneous feedback in the context of free-form responses graded by AI to the best of our knowledge. Improvements to fix grading mistakes in AI involve retraining with more data, improving the model architecture, etc., which may not be feasible due to resource and technological limitations. Furthermore, students hold different attitudes towards AI grading of free-form responses compared to the more hard-coded grading of multiple-choice and programming questions [3, 33]. Therefore, there is a need to understand the learning impacts and to improve workflows and interface designs to help students maximize learning on free-form responses despite inaccurate computer-provided feedback.

The current study contributes to this understanding by using the context of an automated short-answer grader (ASAG) deployed in an introductory computer science course at a large public university in the US. We invited students to participate in surveys and interviews, during which they worked on a series of EiPE problems and got feedback from the autograder as they would in class (Figure 1). The participants then answered a second series of problems composed of slightly modified problems from the first series. We then used causal inference to determine how grading mistakes during the first series of problems influenced performance during the second series. We separately analyzed the impacts of wrong responses marked as right (FPs) and right responses marked as wrong (FNs). This kind of analysis allows us to inform educators' FP-FN trade-off decisions in their deployment of AI systems.

Furthermore, we investigate the *mechanisms* by which FPs and FNs harm learning in order to provide targeted suggestions for educators and designers to mitigate the learning impacts of autograder errors on students. Our causal modeling focuses on several hypotheses that prior work does not completely explore, including behavioral engagement after receiving the feedback, students' ability to detect grading errors, and students' prior abilities. To complement this, we interviewed students to probe how they interacted with inaccurate feedback. We present the formal research questions in Section 2.

From our study, we found distinct effects of FPs and FNs. FPs harmed learning by preventing students from knowing about the conceptual errors in their answers. This was partially because many participants skimmed or skipped the feedback during FPs, and partially because being marked right appeared to bias participants against admitting they were wrong and thus correcting their answers. Higher-performing students detected FPs at a lower rate and experienced a larger absolute drop in their learning outcomes than lower-performing students, possibly because they were more confident in their incorrect answers.

FNs affected our participants less consistently. FNs caused statistically significantly lower learning outcomes in the survey setting but not in interviews. When learning harms did happen, it was often because participants reworded post-test responses in a way that made them too vague to be considered correct. Participants' accuracy in assessing their answers as correct after receiving feedback did not explain the difference between the interview and survey conditions. We hypothesize the additional cognitive engagement that interviewees experienced when explaining their thought processes upon receiving the feedback protected them from the learning harms.

In addition to our findings, we make the following contributions to the design and deployment of ASAGs, especially for EiPE problems:

- Recommendations on workflows and interfaces to help learners detect FPs and enhance effective engagement with feedback when faced with FPs and FNs.
- Empirically-supported advice for determining FP-FN trade-offs in AI-generated feedback.
- Evidence of the value of personalizing AI error mitigation strategies based on student performance level.

2 RELATED WORK

In this section, we first briefly review developments in feedback for free-form responses, including Explain in Plain English problems in computer science. We then discuss feedback in general and the impact of erroneous feedback, especially erroneous feedback from computers. Lastly, we discuss prior work that hints at how autograder errors impact learning.

2.1 Automatic Assessment of Free-Form Responses

Educators are incorporating NLP approaches in their free-form response feedback workflows to offer timely feedback to students at scale. For instance, Automated Short Answer Grading (ASAG) systems evaluate the objective correctness of short answers [14, 43]. They have been deployed in a variety of classroom exercises and high-stakes examinations, including in computer science [3], biology [2], and physics [38] courses. Beyond use in individual classrooms, C-rater, an ASAG system, was used in a National Assessment for Educational Progress (NAEP) assessment and a statewide assessment in Indiana [43].

2.2 Explain in Plain English (EiPE) problems

The ASAG system that we evaluated in this study was deployed in a university introductory computer science class and graded students' responses to problems that asked them to provide short English descriptions of Python code (Figure 1). These problems are known as Explain in Plain English (EiPE) problems. EiPE rose in prominence as a method to assess students' ability to read code and discuss its behavior at a high level of abstraction [67]. While a full discussion of the teaching effectiveness of EiPE in computer science education is beyond the scope of this paper, multiple studies have supported code reading as a necessary skill for developing more advanced programming proficiency [19, 44, 45, 53, 66].

2.3 Formative Feedback

Shute defines formative feedback as information communicated to the learner with the intent to modify their thinking or behavior to improve learning [60]. Feedback is a critical facilitator of learning and performance [4, 5, 20, 34, 60]. A wide variety of factors impact feedback effectiveness, such as specificity [6, 20, 57, 69], complexity [42, 58, 62], timing [1, 18, 48, 59], and more. These factors interact

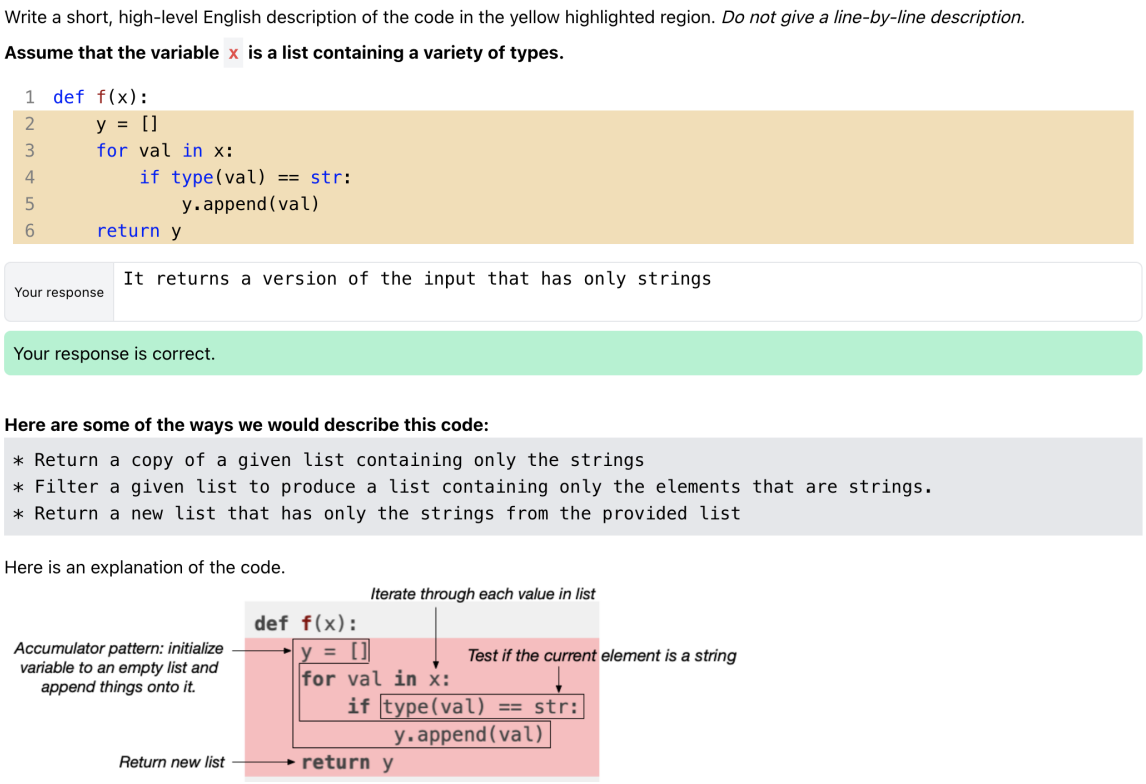

Figure 1: Screenshot of the interface that participants used in our study. The workflow and language mirrors that which was used in class. Students are instructed to provide a high-level description of Python code. Shown is a problem used in the study, along with a participant's response and the feedback the computer provided. The bottom part of the screenshot contains the feedback, including the autograder's mark, exemplar answers, and a labeled diagram. This feedback always appears, regardless of whether the response is marked correct.

with the characteristics of the feedback receiver. For example, lower-performing students benefit more from immediate feedback than delayed feedback [23, 47], and they benefit more from feedback that explains why an answer is right/wrong than feedback that simply states whether an answer is right/wrong [17, 30].

In addition, learners must receive feedback mindfully for it to benefit learning [6, 70]. Some learners engage with feedback well [31], and others poorly [26, 37, 61]. Many personal characteristics of the feedback receiver affect engagement, including gender [64], self-efficacy [29], motivational beliefs [63], and more. Furthermore, feedback receivers need to view a source as credible before they are willing to use the source's feedback [13, 21, 40]. In this paper, we extend the rich body of work on factors that impact feedback effectiveness and engagement by studying the effects of erroneous feedback from AI.

2.4 Impacts of Inaccurate Feedback

A limited number of studies have examined how inaccurate feedback impacts feedback usage and learning. Johnson et al. found that contingent (i.e., accurate) performance feedback improved performance relative to independent feedback uncorrelated with their performance [36]. Brand et al. found that people who received inaccurate verification feedback on simple match-to-sample tasks were more likely to fail to acquire task-relevant skills [12]. This was true both for human- and computer-provided feedback. In addition, inaccurate feedback had a lasting effect in [32], as participants didn't learn with inaccurate feedback and didn't begin learning immediately after switching to accurate feedback.

Our current study adds three novel perspectives. First, we evaluate the impact of inaccurate feedback on AI-graded free-form responses, specifically in the domain of ASAG. Prior work suggests that students hold different attitudes towards AI grading of free-form responses compared to the grading of questions that are more hard-coded, such as multiple-choice and programming questions [3, 33]. Second, in this context, we separately analyze the impacts of wrong answers graded as correct (FPs) and correct answers graded as wrong (FNs). As AI systems designers must make trade-offs between FPs and FNs, a better understanding of the relative harm between these two types of errors will maximize student learning from these systems. Third, as far as we know, we are the first to model how prior student performance or knowledge interacts with erroneous feedback, especially in an ASAG context. This knowledge

will help design personalized interventions relevant to imperfect autograder feedback. Hence, we ask:

- **RQ1: What is the learning impact of FPs and FNs during autograded EiPE problems?**

2.5 How Autograder Errors Impact Learning

Gaining a deeper understanding of *how* autograder errors impact learning will be critical for addressing the impacts of these errors. While there may be many reasons, we focus on a few hypotheses hinted by prior work.

2.5.1 Students' Assessment of Their Answers' Correctness. Prior work suggests that students blame the autograder for marking them wrong and have trouble distinguishing errors from non-errors. We hypothesize not detecting FPs could prevent students from trying to improve their incorrect answers, and hence negatively impact learning. In an investigation of a deployed ASAG system in a computer science course, Azad et al. showed that students perceived the ASAG as less reliable than other types of grading (e.g., multiple choice questions), and students often appealed the autograder's mark even when their answer was truly incorrect. In a later study exploring a similar ASAG system, Hsu and Li et al. found that students overestimated the probability the grader would mark correct answers as incorrect and that many students were unaware that the system marked incorrect answers as correct [33]. This suggests that students may fail to detect whether the autograder is making an error and subsequently inaccurately assess their answers' correctness. In contrast to prior work, we directly probe students' thinking after receiving an ASAG's feedback as they assess whether they wrote a correct response. We further link the detection of autograder errors to learning outcomes.

- **RQ2a: How does students' ability to assess their answers' correctness upon receiving feedback change with autograder errors (FPs, FNs)?**
- **RQ2b: How does accurately detecting or not detecting autograder errors explain learning outcomes?**

2.5.2 Feedback Engagement. Errors in computer-provided feedback may reduce student engagement with the feedback, negatively impacting learning. We expect students will pay less attention to feedback when marked right, even when it happens accidentally. A small usability study for an ASAG in the subject of natural science (N=6, 13 questions per participant) observed that one of the participants failed to notice a FP primarily because they ignored the detailed feedback when marked right [38].

The current study investigates students' behavioral engagement with imperfect ASAG feedback in a larger sample and examines the downstream impacts on detecting autograder errors and learning. We additionally model students' engagement with FNs, which prior work does not address. We formalize the research questions as follows:

- **RQ3a: How does engagement with feedback change with grading errors (FPs, FNs)?**
- **RQ3b: How does this engagement influence students' ability to assess their answers' correctness upon receiving feedback?**
- **RQ3c: How does this engagement explain learning outcomes?**

2.5.3 Prior Knowledge and Experience. As mentioned in Section 2.3, lower-performing students learn differently from feedback compared to higher-performing students. As far as we know, we are the first to model how prior student performance or knowledge moderates the impacts of erroneous feedback, especially in an ASAG context. This knowledge will help us design personalized interventions relevant to imperfect autograder feedback. We reasoned that since low-performers have more to learn and improve than high-performers, inaccurate feedback would lead to more missed learning opportunities for lower-performers. More formally, we check how class performance interacts with autograder errors to produce learning effects, as well as our proposed explanations for these effects:

- **RQ4: How does student performance moderate autograder errors' impact on (1) their learning, (2) accurate assessment of their own answers' correctness, and (3) their engagement on feedback?**

3 CAUSAL MODELING

Our RQs suggest various causal relationships among the autograder's Decision Class (TP/TN/FP/FN), student performance, and three key outcome measurements: engagement time on feedback, accurately assessing the correctness of one's answer, and learning outcomes. Figure 2 visually organizes these relationships into a directed acyclic graph (DAG) according to our RQs.

Unfortunately, we have no practical way to manipulate these factors in an ecologically-valid way. For instance, to assign a student to experience a FP (a wrong answer graded as right), we would have to force them to write a wrong answer. Likewise, manipulating whether a student detects an autograder error requires interfering with their natural cognitive processes.

We instead take a quasi-experimental approach that observes behaviors and controls for confounds. Prior to collecting any data, we performed a literature review and identified confounding factors – factors we reasonably believed could each correlate with more than one variable of interest (the variables on the right side of Figure 2). We grouped confounds based on the subset of variables that they affected:

- Moderating confounders – student performance and our experimental setting (survey or interview). These moderate all the relationships on the right side of Figure 2.
- Group 1 confounders – factors that could predict the likelihood of a participant writing a correct response as well as change their usage of feedback, and thus have the potential to correlate with both Decision Class and our outcome measurements. We included participant ID, question ID, question order (due to fatigue [56] or learning from earlier questions), and a participant's confidence in a question [52].
- Group 2 confounders – we judged these to potentially correlate with our three outcome measurements but not with Decision Class. They include task-value belief, task success expectancy, prior trust of the autograder, and more. For a full list and justifications, please see Appendix 9.1.

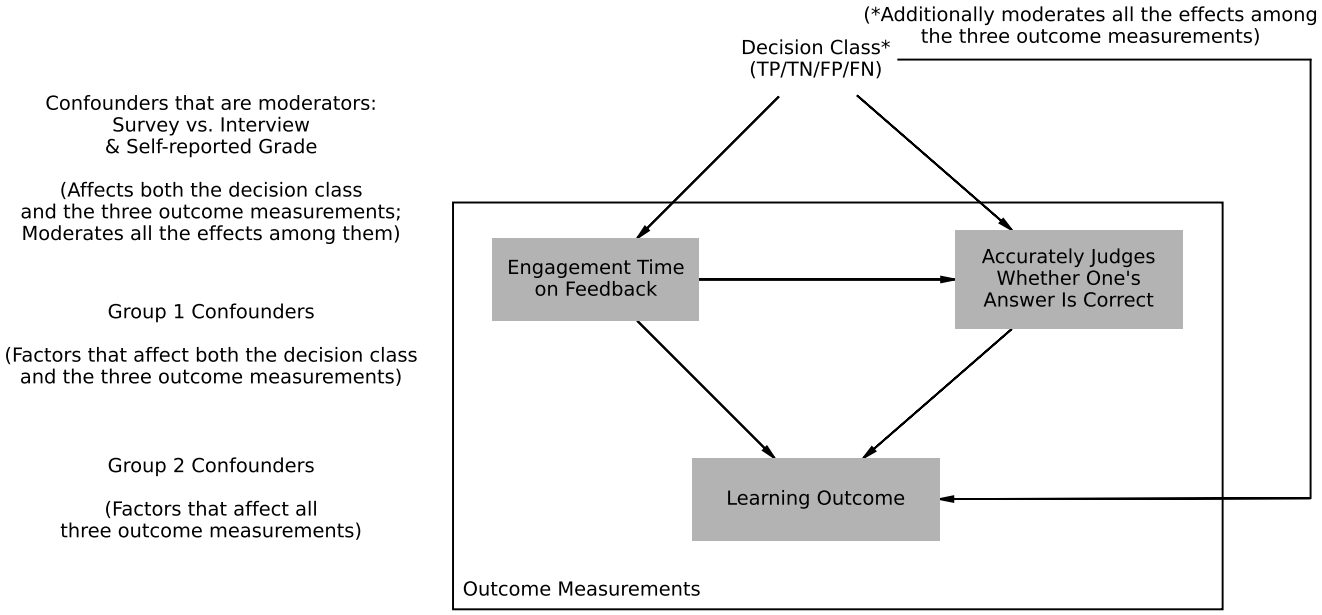

Figure 2: Our main causal diagram describing effects in the context of a single instance of an autograded EiPE question. Edges in this directed acyclic graph (DAG) represent researchers' belief that one variable may affect another based on prior work and reasoning. The left side of the figure summarizes confounds. Edges implicitly exist between the confounds and variables on the right side of the figure as described but are not drawn to reduce visual clutter.

This grouping provides specificity about which confounds affect which variables, which helps avoid needless inclusion of covariates and careless introduction of confounds that could cause spurious correlations during statistical modeling [49]. We incorporated these confounds into the DAG (left side of Figure 2), then derived several regressions to answer our RQs. Table 1 lists these regressions. Each regression tests a subset of the DAG; taken together, this organization is conceptually similar to Baron and Kenny's method of using several regressions to test statistical mediation [8]. Models 1, 2, and 3 examine the total effects of Decision Class on outcome measures; they include but do not separate out the effects via intermediary variables. Models 4 and 5, in contrast, do measure the effects of intermediary variables separately. Supplementary materials include the detailed specifications of the models.

4 METHODS

To answer our research questions, we conducted a quasi-experiment where we asked students to work on a series of autograded EiPE problems in either a survey or an interview environment. In this section, we first describe the context of the study and the background of the autograder. Then we describe how we operationalized the measurement of participants' engagement time, their assessments of their answers' correctness, and learning during the interviews and surveys. Next, we discuss the ethical considerations of our study design. Finally, we describe the participant recruitment, data annotation, and analysis approaches.

4.1 The class setting and autograded questions

This study took place in the context of a full-semester introductory computer science class for non-majors at the University of Illinois at Urbana-Champaign in Fall 2021. Approximately 600 students were enrolled in this class designed to teach Python and Excel to people without prior programming experience. Due to the COVID-19 pandemic, the course's lectures were conducted synchronously online. Even before the pandemic, all homework and exams were computerized.

4.1.1 Development and deployment of the autograder. The course mentioned above implemented EiPE problems with an autograder that marked answers right or wrong, allotting no partial credit. This autograder was developed specifically for the course, and fit a logistic regression model on bigram and bag-of-words features split and tokenized with Python's *nltk* module. It was approximately 87% accurate, which was statistically indistinguishable from the course's TAs accuracy [22]. Problems were presented to students as shown in Figure 1.

The autograder graded both formative assessments (homework) and low-stakes summative assessments (four small quizzes worth 2% each). On homework, the system provided feedback in the form of the autograder's mark, exemplar answers, and sometimes a labeled diagram of the code for more complex problems (see Figure 1).

In multiple lectures at various times throughout the semester, the instructor formally discussed the autograded EiPE questions. They explained that answers should be unambiguous, correct, and

Table 1: Piece-wise modeling of the causal diagram: five regression models. We use the following abbreviations: IVs - Independent Variables; DV - Dependent Variable or the response variable of the regression; CFs - relevant confounders, which are controlled for by adding them as additional predictors for each DV. In all the models, for each IV, we additionally include an interaction effect between that IV and the experiment setting (survey vs. interview); this is not explicitly written inside the table for purposes of brevity. Note that while we believe self-reported grade moderates all effects, we only added this interaction effect to the Decision Class variable in the total effect models (Model 1-3) and the Engagement Time and Accurately Assessed Own Answer variables in Model 5, in an effort to increase parsimony.

#	IVs	DV	CFs	Comments
1	Decision Class (D), Self-Reported Grade (G), D x G	Learning Outcome	Group 1	Addresses RQ1 (total effect of autograder errors on learning) and RQ4 (how the impact differs with student performance)
2	Decision Class (D), Self-Reported Grade (G), D x G	Accurately Assessed Own Answer	Group 1	Addresses RQ2a (total effect of autograder errors on whether learners accurately assess their answers' correctness) and RQ4 (how they differ with student performance level)
3	Decision Class (D), Self-Reported Grade (G), D x G	Engagement Time	Group 1	Addresses RQ3a (total effect of autograder errors on engagement) and RQ4 (how the impact differs with student performance)
4	Decision Class (D), Engagement Time (E), Self-Reported Grade (G), D x E	Accurately Assessed Own Answer	Group 1, Group 2	Addresses RQ3b (direct effects of engagement and autograder errors on whether learners accurately assess their answers' correctness)
5	Decision Class (D), Engagement Time (E), Accurately Assessed Own Answer (A), Self-Reported Grade (G), D x E, D x A, G x E, G x A	Learning Outcome	Group 1, Group 2	Addresses RQ2b, RQ3c (downstream learning effects of engagement, detecting errors) and RQ4 (how the impact differs with student performance)

at a high level of abstraction in order to receive credit. They further presented and explained authentic student errors and explained that most perceived issues were not autograder errors. Students were then given some sample answers to these problems to grade for themselves.

4.2 Approach 1: Interview

We conducted 30 semi-structured interviews to gather qualitative data on interactions with autograded EiPE problems in addition to quantitative data. Interviews took place over Zoom while participants simultaneously completed an interactive Qualtrics questionnaire. The questionnaire embedded an EiPE interface similar to the interface that participants were using in class (Figure 1). Interviews lasted 80 minutes on average, and participants were paid $15.00 per hour, with up to $2.00 in bonus compensation (explained later).

The first two co-authors both functioned as interviewers. To ensure consistency, the interviewers co-conducted the 1st, 2nd, and 12th interviews. Besides those three, each interviewer conducted 12 and 15 interviews solo, alternating as much as possible within the constraints of each person's schedule. As new situations appeared, the interviewers communicated their experiences and proposed refinements to the wording of follow-up questions. Minor revisions happened after the 3rd, 5th, and 12th interviews. We summarize the interview protocol below; the supplementary material contains all details and exact question wording.

Part 1: Beliefs Towards EiPE Questions and the Autograder. First, we elicited participants' perceptions of the autograder's error rates (a variable in Group 2 confounders, CF2), how valuable they thought EiPE problems were (CF2), and how well participants expected themselves to do on EiPE problems (CF2).

Part 2: Practice EiPE Questions with Autograder Feedback (Pre-Test). Next, participants encountered eight EiPE problems in a randomized order. For each problem, participants also rated their confidence in understanding the code (a variable in Group 1 confounders, CF1) and confidence that the autograder would grade their answer as correct.

Upon submitting an answer to each problem, participants received feedback in the same manner as the course's homework assignments. The feedback included the autograder's judgment of correctness, some exemplar answers, and a code explanation diagram. However, not all eight questions were graded by the same autograder used in class. To increase encounters with FPs and FNs, two problems were always marked as correct and two always as

incorrect – for some participants, this was the *only* way they experienced errors. Participants were not informed of this change until the debrief at the conclusion of the study. We tracked the number of milliseconds between the time the feedback appeared and the time the participant clicked "Next" to move on to the next question, which operationalized our Engagement Time outcome measurement.

Next, we collected participants' assessments of their responses' correctness two times. The first time, we hid the problem statement and feedback to reduce participants' tendencies to reflect substantially on it. We used this first opinion during quantitative analysis, as we considered it more ecologically valid, closer to the opinion that participants would have in a natural homework environment. The second time, we restored the problem and feedback to allow a richer discussion and for participants to walk through what they did with the feedback.

Part 3: General Feedback Usage Habits. Here we asked about participants' typical habits using the feedback on their homework and how their behavior during the interview differed.

Part 4: EiPE Post-Test. The second block of EiPE questions showed perturbed versions of the problems in the first block in order to measure question-level learning outcomes. Questions appeared in a randomized order independent of the first block. To increase the stakes and encourage maximal effort, we told participants that each correct answer in this block would earn them $0.25 in bonus compensation. We also told participants not to use outside resources. No autograding happened in this phase, and each question directly proceeded to the next one without feedback.

Part 5: Personal Characteristics and Demographics. We surveyed participants' expected grades in the class (measuring student performance level), goal orientation (CF2), and demographics (CF2), including gender, English proficiency, etc.

Part 6: Debrief. Finally, we told participants that some of the problems were, in fact, not graded by the autograder and were always marked as right or wrong. We re-displayed the questions from the first block, and highlighted the statically-graded questions. The interviewer asked participants to review all the questions and immediately provided manual grading and explanations to help participants understand whether their responses were actually correct or not. We provided the same kind of feedback on their post-test answers via email within a week.

4.3 Approach 2: Online survey

To gather more data, and in an environment more similar to that of the homework without additional induced reflection, we modified our Qualtrics questionnaire to collect data without an interviewer. Participants were compensated $10 for completing the survey, with up to $1.50 in bonus compensation. The median completion time was 22.8 minutes, corresponding to a median rate of $26.3/hr before bonuses. We made the following modifications to the questionnaire:

- During the first series of pre-test EiPE problems (Part 2), we cut two normally-autograded problems to decrease survey fatigue. That left us with two normally-autograded problems, two always-marked-correct problems, and two always-marked-incorrect problems. In addition, we asked no additional survey questions in between problems to avoid distractions and reflection in excess of a regular homework environment. To measure whether participants thought they had written a correct response, we displayed a button next to the feedback that said "I think the autograder may have made a mistake" to allow reporting.
- At the end of Part 4, we asked participants if they had used outside resources during the post-test and told them that their response would not impact their compensation. One participant reported that they did, and their data was discarded.
- We asked survey participants if they had not reported autograder mistakes when they in fact thought they occurred. Participants who indicated *Yes* were excluded from the three models that utilized the variable of assessing answers accurately (Models 2, 4, and 5) because their reporting behavior did not accurately represent whether they thought their response was correct. We are confident such exclusion did not introduce biases or confounds because we controlled for factors associated with participants not reporting autograder errors, e.g., self-reported grades, prior trust of the autograder, and confidence in the question.
- During the debrief (Part 6), while we did re-display the problems that were always graded as right or wrong, there was nobody to provide immediate manual grading or comments. We instead encouraged them to reach out to us if they had immediate questions and emailed manual grades and comments to participants within one week after their survey completion.

4.4 Constructing the pool of EiPE questions

By the time of our study, students were already exposed to and using the autograder in their course. To reduce the chance of students memorizing answers from existing assessments for use in the interviews and surveys, we created a pool of new EiPE questions. The question pool was designed to have three features. First, the questions had to support easy modification so we could create a new question for the same concept on a post-test, and thus a one-to-one correspondence between pre-test and post-test questions. Second, we had questions cover non-overlapping concepts as much as possible to minimize the learning effect across pre-test questions. Third, we produced a range of difficulties to show different feedback usage relative to question difficulty. A course instructor created eight pre-test questions and perturbed all eight into modified versions for use on the post-test. Seven out of eight modified post-test questions had their operators inverted, i.e., switching instances of "<" to ">", "+=" to "-=", etc. The remaining question had an if-even statement switched to an if-odd statement; that is, from "if n % 2 == 0" to "if n % 2 == 1". All questions are available in the supplementary materials.

4.5 Ethical considerations

As previously mentioned, setting some questions always to be marked as correct or incorrect was the only way for some participants to experience a FP or a FN. We carefully considered the risk of negative learning and emotional effects from autograder errors in our study. In summary, both we and our institution's IRB judged that the study would not cause harm or distress exceeding what students in the class already encountered. Factors that mitigated the risk of emotional harm included the knowledge that students had already been experiencing autograder errors in their regular homework assignments and that the study was worth no credit. The debriefing process mitigated concerns of learning harm. We acknowledge that our design did not guarantee survey participants' engagement with the debriefing; however, we felt that ongoing participation in the course would mitigate harm for such students.

4.6 Participant Recruitment

Via email announcements, we recruited participants from the university introductory computer science course from November 9, 2021 to November 29, 2021 – the latter part of the semester, but before final exams. We distributed announcements for the interview first, expecting interviewees to be more difficult to find. Students could only participate in either the survey or the interview; announcements and the survey landing page clarified this, and we checked for duplicate contact emails.

Based on our prior experiences, we expected an overrepresentation of higher-performing students among those volunteering to participate. To try to recruit a more balanced population, students in the lowest 50% percentile of grades in the class received more reminder emails about the study than the rest of the class.

We collected 40 complete survey responses and conducted 30 interviews. Due to reasons such as failing attention checks, referencing outside sources, and obvious misunderstanding of key questions, we excluded 4 survey responses and 1 interview from the analysis. Of the remaining participants, there were 216 question pairs from surveys and 232 from interviews. Table 2 presents the descriptive statistics of the dependent and independent variables in our full sample. Note that among the survey participants, 17 out of 36 said they did not report all autograder errors that they detected, and their data were excluded from models 2, 4, and 5. We still had 114 question pairs from the survey after the exclusion.

38% and 47% of the participants self-identified as male in interviews and surveys, respectively, and the rest as female. Around 68% of participants reported "A" as their expected grade in both interviews and surveys. Because of the small numbers of "B" and "C" participants (and no other values in our data set), we chose to group them together. This is close to the distribution of the entire class, where half of the students received a final grade of "A". Based on this grouping, we created the binary class performance variable with values "higher-performers" and "lower-performers".

4.7 Data Annotation

Prior to analysis, we annotated various data, including the "ground truth" decision class of all responses to the EiPE problems in the interview and survey, and interviewees' first impressions of whether they thought their response was correct. In each annotation task, two team members independently and deductively coded the data, achieving high inter-rater reliability (IRR), and then a third team member functioned as a tie-breaker. For more details about the annotation process and IRRs, please see the supplementary materials.

4.8 Bayesian Modeling

We performed the five regressions laid out in Table 1 with Bayesian formulations of hierarchical generalized linear models. All models made predictions on the participant-question level – in other words, each "data point" was one participant's behavior on one practice question and its corresponding post-test question. We defined learning outcomes in Models 1 and 5 as a binary outcome of whether a participant answered a post-test question correctly, regardless of their correctness on the practice version of the same question. Models 2 and 4 predicted the binary outcome of whether a participant assessed their answer's correctness accurately. The models with binary outcomes – all of them except Model 3 – utilized a Bayesian logistic regression.[1] Model 3 performed a more standard Bayesian linear regression predicting the log-transformed feedback engagement time data, which allowed us to describe proportional instead of absolute differences in engagement time.[2]

Bayesian modeling requires priors, which embody prior beliefs and evidence about the distribution of modeled variables and effect sizes. We followed two general principles to set weakly informative priors. First, we selected maximum entropy distributions (e.g., uniform, normal, exponential distributions), which innately make the most conservative inference given the parameters and the data [49]. Second, we set the parameters of these prior distributions to encode skepticism but not impossibility towards large effect sizes. For instance, our priors considered an autograder error increasing time spent on feedback by tenfold an unlikely outcome. This increased statistical power, useful for our relatively small sample size, and ensured the priors did not dominate the findings.[3] The full model specifications can be found in the supplementary materials.

Given the priors and observed data, Bayesian regression generates a set of posterior distributions representing the likely values of regression coefficients. In general, no closed-form solution describes the shape of the posterior distribution, so implementations of Bayesian inference use sampling strategies. To obtain our posterior distributions, we implemented the models with NumPyro, a popular Bayesian inference framework, and performed sampling using the Markov Chain Monte Carlo (MCMC) technique with the No-U Turn Sampler (NUTS). For all the models, all parameters achieved a Gelman-Rubin statistic (a measure of MCMC convergence) of 1.0, indicating that the multiple sampling chains converged [25].

[1] More precisely, we model the data as a Binomial distribution and the distribution's probability parameter as a logistic regression.

[2] A log transform also helps ensure additivity and linearity among predictors [24]. Model 3 models the transformed data as a normal distribution and the distribution's mean parameter as a linear regression.

[3] Besides the advantages gained in the way we set priors, Bayesian modeling also provides transparency about the model's assumptions and the ability for future researchers to incorporate our findings into their priors. Kay et al. further elaborate on the advantages of Bayesian inference in HCI in [39]. König and van de Schoot similarly advocate for a Bayesian approach in the educational research context in [41].

Table 2: Descriptive statistics of the dependent variables and independent variables by experiment setting and decision class in our full sample. Note that the proportion of FPs and FNs in the sample is inflated compared to that of the autograder used in the course.

Setting	Interview				Survey			
Decision Class	**TP**	**TN**	**FP**	**FN**	**TP**	**TN**	**FP**	**FN**
# Instances	51	94	50	37	57	88	52	19
% Accurately assessed correctness of own answer	98.04	68.09	32.00	78.37	100.00	77.27	3.85	52.63
% Correct on post-test	78.43	37.23	28.00	72.97	77.19	31.82	9.62	31.58
Seconds on feedback [Median]	19.41	26.31	25.92	31.37	4.17	8.78	6.77	14.35
Perceived FPR (%) [Mean (SD)]	11.31 (9.98)				10.47 (8.34)			
Perceived FNR (%) [Mean (SD)]	36.93 (25.62)				29.31 (24.10)			

4.9 Thematic extraction from interviews and open-ended survey responses

In general, we used an iterative open coding process to inductively generate themes that had relevance to the three outcome measures of the RQs – engagement with feedback, ability to assess the autograder's feedback accurately, and learning. We examined four types of qualitative data: (1) open-ended survey responses about participants' usual habits on the feedback, (2) general themes from the interview transcripts, (3) reasons why interviewees disagreed with the autograder's decision (61 instances), and (4) paired pre- and post-test responses.

The exact analysis process varied by the type of data and questions we wanted to answer. Overall, for data types 1, 2, and 3, two or more team members conducted the coding process and created a separate codebook for each data type. For data type 4, a course instructor performed the analysis. For more details, please see the supplementary materials.

5 RESULTS

In this section, we first introduce how to interpret our Bayesian model results. Then, we present the learning impacts of FPs and explanations for the effects, followed by the same for FNs.

5.1 Effect Sizes and Statistical Significance

We describe the effect of autograder errors (RQ1, 2a, 3a) by reporting our models' predictions of *what would have happened if the autograder had graded a problem instance correctly*. Namely, this compares FPs to TNs and FNs to TPs. As an example, we describe the effect of FPs on learning (RQ1) compared to the effect of TNs. In this paper, we express all effects on binary outcomes as an odds ratio [11], so an effect size of 2 means that participants were twice as likely to answer correctly on the post-test after FPs compared to TNs; 0.5 means half as likely, etc.

We discuss the relationships among feedback engagement time, accurate answer assessment, and class performance (RQs 2b, 3b, and 3c) slightly differently. We report effects among these three outcome measurements *given that a FP or FN happened*, which reflects our interest in the downstream effects of autograder errors. Conditional on FP or FN, effects can be interpreted similarly to traditional regression coefficients: the relative (proportional) change in a DV given a unit change in the IV. The unit of change for engagement time is the sample standard deviation of $log(time\_spent)$. A unit change in the remaining binary variables (assessed answer correctly, student performance) means switching from one level to the other. For example, given a FP, we examined whether successfully recognizing one's own response as wrong influenced learning outcomes. Again using an odds ratio, an effect size of 2 means that compared to not detecting the FP, detecting the FP doubled the odds of writing a correct post-test answer.

For each effect, our Bayesian formulation provides a posterior distribution expressing a range of likely effect sizes. We report the mean of this distribution and the boundaries of the 94% Highest Posterior Density Interval (HPDI) – the narrowest interval containing 94% of the probability mass.[4] We declare statistical significance (i.e., a causal effect exists) when the entire HPDI lies outside of the reference value of 1 – that is, when the proportional change in DV caused by a unit change of an IV is highly likely to be different than 1.

5.2 Total effect of FPs on Learning (RQ1)

Marking an incorrect answer as right reduced the likelihood of a correct post-test answer. Out of all four decision classes, Model 1 predicts that participants who experienced a FP during a practice question were the least likely to answer a perturbed version of the same question correctly later (Figure 3). Erroneously marking an incorrect answer as right instead of wrong (FP vs. TN) while

[4]We selected 94% as a relatively strict yet arbitrary norm, just as how p-value < 0.05 is arbitrary.

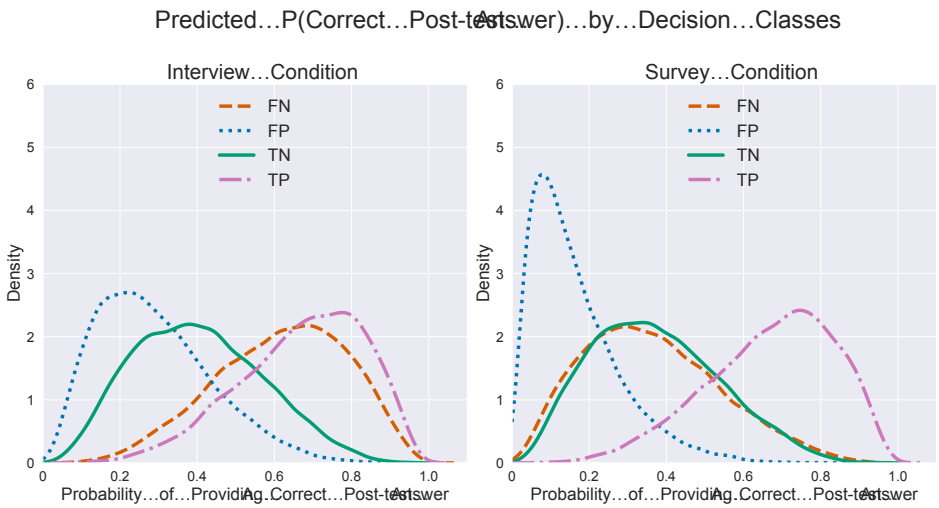

Figure 3: *Main findings:* Participants were worst at getting the post-test question right after experiencing FPs, regardless of survey or interview. Given a correct answer during the practice session, being erroneously marked wrong (FN) predicted worse learning outcomes compared to being marked right (TP). But on the interview, the outcomes between FNs and TPs were about the same. *Reading the figures:* Each figure contains the posterior distributions of the predicted probability of a correct post-test response, one for each decision class (FN/FP/TN/TP). Each curve is constructed from 15,000 posterior distribution samples. The more concentrated to the left a distribution is, the lower the inferred learning outcome for that decision class.

holding all else constant decreased the predicted chance of a correct post-test answer from a mean of 40.3% to 28.8% in the interview. The reduction is statistically significant with a small effect size (FP vs. TN odds ratio posterior: M = 0.58, 94% HPDI = [0.41, 0.74])[5]. The effect is larger in the survey with a small to medium effect size (FP vs. TN odds ratio posterior: M = 0.36, 94% HPDI = [0.23, 0.48]), where the drop is from 37.2% to 16.3%.

Qualitatively, participants that wrote an incorrect post-test response after a FP often failed to fix their mistakes from the practice questions. There were 91 such instances of a wrong answer after a FP. In 55 of these instances (60%), the participants either made the same mistake in the post-test as they did during the practice questions or wrote an answer that contained more errors. In another 26% of the instances, the participants wrote improved post-test answers, but failed to fix every mistake. Participants in the remaining cases (14%) wrote post-test answers that would have been correct if they were written for the pre-test question, possibly due to memorization of exemplar answers in the feedback or an assumption that questions in the pre-test and post-test were identical.

5.3 Explanations of FPs' learning harm (RQ2 & RQ3)

We found that FPs harmed learning. Now, we describe how participants' engagement time and ability to detect autograder errors explain this effect. Figure 4 summarizes these explanations, which we detail below.

[5]We interpreted the magnitudes of odds ratios based on Chen et al. [16], where odds ratios of 1.68, 3.47, and 6.71 are equivalent to Cohen's d = 0.2 (small), 0.5 (medium), and 0.8 (large), respectively. If an odds ratio is less than 1, we take the inverse of the odds ratio before assessing its magnitude.

5.3.1 FPs and engagement time (RQ3a). **Participants spent less time on their wrong answers because of FPs**, according to our causal inference from Model 3 (Figure 4, Decision Class → Engagement Time, effect size expressed as ratio of engagement time). Figure 5 displays an intuitive comparison of the absolute effect sizes. In surveys, for example, FPs contributed an average reduction of 1.4 seconds with respect to 4.8 seconds of average time spent on TNs.

We observed interviewees often skipping or skimming feedback when receiving positive feedback from the autograder, and many participants reported they had similar skipping habits on homework. Interviewees also often cited confidence in the response they wrote as a reason for skipping feedback. Of course, skipping feedback was only appropriate if one actually wrote a correct response. These findings suggest that negative feedback would have motivated participants to examine all the feedback, or at least doubt their answers.

5.3.2 FPs and error detection (RQ2a, 3b). The isolation of the FP distribution in Figure 6 shows that participants assessed FPs the least accurately among all decision classes. Given an interviewee wrote an incorrect answer, Model 2 predicts that they acknowledged the mistake with a 73% probability when properly marked wrong (TN) and 37% probability when erroneously marked right (FPs), a drop of 36%. The equivalent drop in the survey is larger, from 75% to 20%. The difference between TNs and FPs is statistically significant with a medium to large effect size (FP vs. TN odds ratio posterior; Interview: M = 0.35, 94% HPDI = [0.24, 0.47]; Survey: M = 0.14, 94% HPDI = [0.07, 0.22]).

Model 4 confirms that **engagement with feedback played an essential role in detecting FPs.** We found a statistically significant

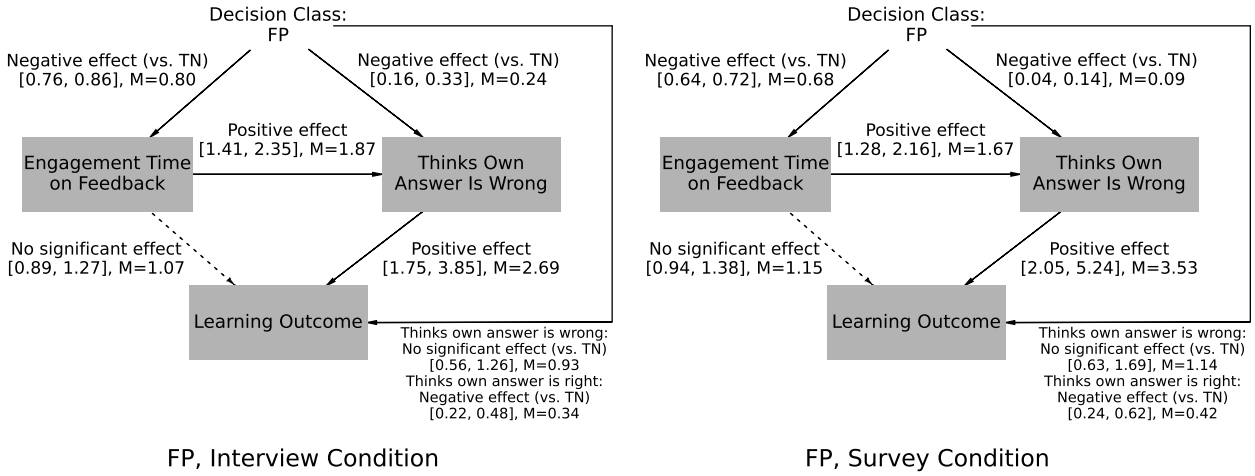

Figure 4: Annotated causal diagrams summarizing effects given an FP and separated by experimental setting (survey vs interview). All the effects in our main DAG are statistically significant except for Engagement Time → Learning Outcome. Each edge describes the posterior distribution of the effect size, including mean ("M=") and the boundaries of the 94% HPDI (numbers in brackets) – see Section 5.1 for a detailed explanation. A solid arrow indicates a statistically significant effect; a dashed arrow indicates no statistically significant effect. Between survey and interview, effect sizes differ but statistical significances remain the same.

positive relationship between engagement time and detecting FPs with a small effect size (Figure 4, Engagement Time → Thinks Own Answer Is Wrong). In other words, engaging for a shorter duration lowered the odds of assessing one's answer accurately when encountering a FP. As an example, P22 noted during one problem where they received a positive mark that *usually I don't read the correct answer examples. ... If I hadn't read [the exemplars], I would have been like, "Okay, sweet, I'm right." ... I would have not realized [that I was wrong].*

In addition, **FPs directly lowered the probability for participants to admit their incorrectness** with a medium to large effect size (Figure 4, Decision Class → Thinks Own Answer Is Wrong). This suggests that FPs biased participants against recognizing their errors, even when a FP did not change the duration of feedback engagement. Indeed, we observed many instances of interviewees agreeing with the autograder during FPs even after reading the feedback, which represented missed learning opportunities. In one category of agreement, participants claimed conceptually incorrect answers matched the exemplars. Many simply said their response was "similar" to the exemplars, suggesting a focus on superficial similarities while having a very limited understanding of the exemplar answers. In another type of agreement, participants wrote conceptually correct answers that contained too much ambiguity to be considered fully correct. These participants claimed their responses were similar to the exemplars' meaning or essence, but these claims implicitly or explicitly carried ignorance, neglect, or disagreements with the class standards surrounding the use of clear, high-level language.

5.3.3 Learning effects of engagement time and accurate answer assessment in the context of FPs (RQ2b, 3c). **Detection of FPs enhanced learning outcomes.** Admitting that one's incorrect response was wrong given a FP had a small to medium positive effect on learning outcomes (Figure 4, Thinks Own Answer Is Wrong → Learning Outcome). But given that one failed to realize mistakes in their response, FPs decreased learning outcomes more than TNs with a small to medium effect size (Figure 4, Decision Class → Learning Outcome). In other words, FPs not only decreased the likelihood of detecting mistakes in one's answer, but also worsened the consequences of failing the detection.

Engagement time impacted learning outcomes by influencing the ability to detect errors. Given a FP happened, we found no statistically significant direct link between engagement time and learning outcomes (Figure 4, Engagement Time → Learning Outcome). This leaves only the indirect causal path that Figure 4 illustrates: FP → decreases engagement time → lowers error detection likelihood → worsens learning outcome.

A summary of the mechanism behind FP's learning impact: FP's learning harm occurs through several causal paths simultaneously.

- **Less engagement:** Participants spent significantly less time on FPs than TNs, decreasing their odds of recognizing that their answers were wrong and thus leading to worse performance on the post-test question.
- **Bias against recognizing one's error:** Even controlling for engagement time, participants still had significantly lower odds of admitting that their answer was wrong for FPs than TNs.

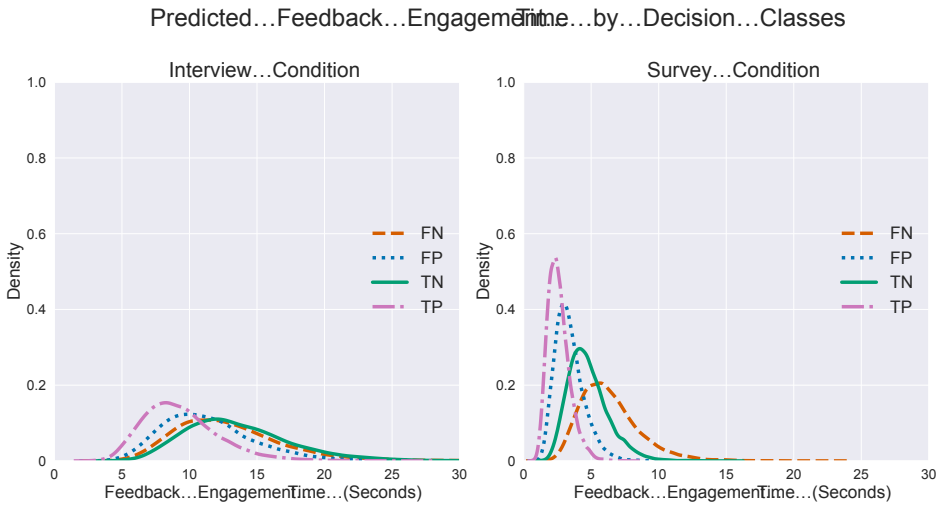

Figure 5: ***Main findings:*** **when participants wrote a wrong answer, erroneously marking them right (FP) decreased engagement time compared to marking them wrong (TN). When they wrote a right answer, erroneously marking them wrong (FN) increased engagement times compared to marking them right (TP). Participants overall spent more time on the feedback in interviews than in surveys.** ***Reading the figures:*** **Each figure contains the posterior distributions of predicted feedback engagement times, one for each decision class (FN/FP/TN/TP). Each curve is constructed from 15,000 posterior distribution samples. The more concentrated to the left a distribution is, the shorter the inferred engagement time for that decision class.**

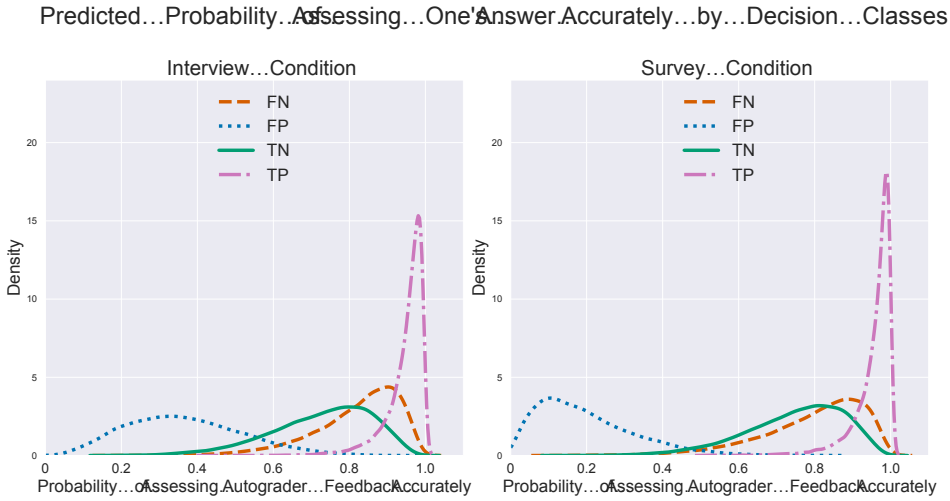

Figure 6: ***Main findings:*** **the distributions representing FPs lie on the left side, indicating participants failed to give a correct assessment of their answers (i.e., recognize their errors) more than any other autograder decision outcome. Participants had a similar probability of accurately assessing FNs and TNs. They very readily accepted TPs.** ***Reading the figures:*** **Each figure contains the posterior distributions of the predicted probability of accurately assessing one's answer's correctness, one for each decision class (FN/FP/TN/TP). Each curve is constructed from 15,000 posterior distribution samples. The more concentrated to the right a distribution is, the better the inferred ability of participants to assess the correctness of their answer for that decision class.**

- **More harmful when failing to recognize one's error:** Failing to acknowledge a response's incorrectness lowered learning outcomes. FPs amplified this effect compared to TNs.

5.4 Total effect of FNs on learning (RQ1)

Previously, we described the effects of FPs, or wrong answers marked as right. Now, we describe FNs. **Marking a correct answer as wrong (FN) reduced learning outcomes in the survey but not in the interviews.** In the interview, Model 1 predicts that participants, on average, answered the post-test question correctly with a 61.7% probability after experiencing FNs (Figure 3), not significantly different from the 67.1% probability after experiencing TPs (FN vs. TP odds ratio posterior: M = 1.01, 94% HPDI = [0.67, 1.37]). But on the survey, the model predicts 36.3% for FNs and 67.0% for TPs. This difference is statistically significant with a small to medium effect size (FN vs. TP odds ratio posterior: M = 0.29, 94% HPDI = [0.18, 0.41]).

Our qualitative observations detail how FNs had a limited impact on post-test performance. During the interviews, when participants explicitly expressed they thought they were wrong when they were actually right, none of them thought they had written a response with conceptual issues. Instead, they mostly identified neutral wording differences between their answers and the exemplar answers that they thought made their answers violate a class standard (e.g., their language was too low-level or contained ambiguity). Out of the 23 instances of FNs in which the post-test answers became incorrect, all except one demonstrated a correct conceptual understanding of the code.

Still, FNs could result in incorrect beliefs surrounding answer construction. One such example was P17, who said *I saw I was wrong, so I read every possible other way I could describe the code ... I realized I didn't have to include "from list x."* However, this directly contradicted the class standards requiring specificity about the inputs to the code, and "from list x" provided exactly this specificity. Later, P17 got the post-test version of the problem wrong because they did not specify a precise input.

5.5 Explanations of FN's learning harm (RQ2 & RQ3)

As a reminder, we found FNs impacted learning only on the survey. Similar to our method for FPs, we examined how engagement time and the ability to detect the autograder's errors explain this effect. Figure 7 summarizes our findings, which we detail below.

5.5.1 Ability to detect FNs and learning impact (RQ2a, 2b). **FNs often caused participants to conclude their correct answers were wrong, but with no significant learning impact.** Compared to TPs, FNs overall reduced the odds of concluding one's correct answer was correct with a medium to large effect size (Figure 6; FN vs. TP odds ratio posterior; Interview: M = 0.13, 94% HPDI = [0.08, 0.20]; Survey: M = 0.11, 94% HPDI = [0.05, 0.19]). However, Model 5 found no statistically significant relationship between detecting FNs and post-test outcomes (Figure 7, Thinks Own Answer Is Right → Learning Outcome). But as our example in Section 5.4 reveals, failing to detect FNs could still have negative impacts in some cases. We interpret the statistical result as describing a neutral effect on learning outcomes in *most* cases.

5.5.2 FN and engagement time (RQ3a, 3b, 3c). **FNs increased engagement time, but engagement time had no statistically significant learning impact.** Compared to TPs, FNs increased predicted average engagement time by 133% on the survey and 34% on the interviews (Figure 7, Decision Class → Engagement Time, effect size expressed as ratio of engagement time). Participants spent a similar amount of time on FNs as TNs (Figure 5) – in other words, participants tended to spend time on FNs as if their correct answers were actually wrong.[6]

Yet, we found no statistically significant link between engagement time and learning outcome in the context of FNs (Figure 7, Engagement Time → Learning Outcome). We did find a significant link between engagement time and the ability to accurately assess one's answer as correct in the interviews (Figure 7, Engagement Time → Thinks Own Answer Is Right), but as discussed above, detecting FNs had no statistical relationship to learning outcomes.

5.5.3 A direct negative effect of FNs on learning. According to Model 5, neither engagement time nor error detection ability explain the negative influence of FNs on survey learning outcomes, and thus a "direct" effect of FNs remains (Figure 7, Decision Class → Learning Outcome). The mean of this effect, 0.25, represents a medium effect size. We expand on possible explanations for this direct effect and why it only happened in the survey in our discussion section.

A summary of the mechanism behind FN's learning impact: Neither engagement time with feedback nor the detection of FNs explains the impact. A direct negative effect of FNs on learning in the survey remains.

5.6 How student performance moderates autograder errors' learning harm (RQ4)

In the previous sections, we described the main learning effects of FPs and FNs, and focused on engagement time and the ability to detect autograder errors as explanations. Now, we describe how these learning effects and explanations differed by student performance. This section primarily utilizes Models 1, 2, 3, and 5 (Table 1), which included interaction effects with self-reported grades, a proxy for student performance.

5.6.1 How learning harms of autograder errors differed by class performance. Unsurprisingly, lower self-reported class performance predicted lower post-test scores, regardless of whether the autograder made an error. But whether autograder errors harmed lower-performers' learning more than higher-performers' learning is a different matter. Student performance did not significantly moderate the learning impact of FNs or FPs – except in one case. **On the survey only, FPs harmed higher-performers' learning outcomes slightly more than lower-performers**. Compared to TNs, higher-performers' predicted post-test accuracy dropped

[6] Our model actually predicts a statistically significant difference in time spent between FNs and TNs, but the practical difference is small – 1 to 3 seconds. In the survey setting, this can be explained by the time taken to click the "report" button for FNs.

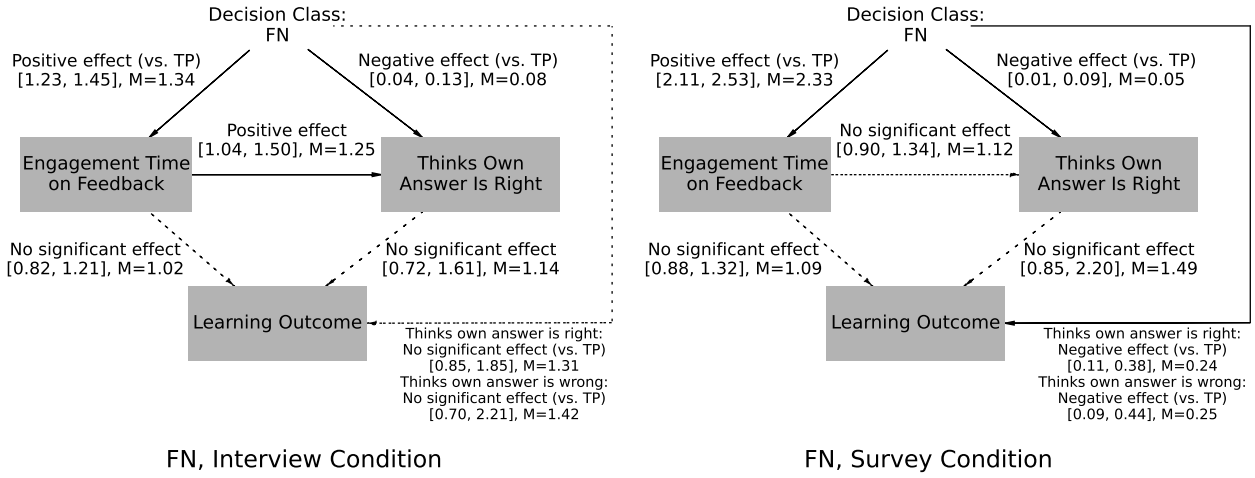

Figure 7: Annotated causal diagrams, describing effects given an FN and separated by experimental setting (survey vs interview). In the interviews, our models found that neither FNs nor any mediating factors significantly affected learning outcomes. In the survey, FNs only affected learning outcomes when participants failed to detect them. Each edge describes the posterior distribution of the effect size, including mean ("M=") and the boundaries of the 94% HPDI (numbers in brackets) – see Section 5.1 for a detailed explanation. A solid arrow indicates a statistically significant effect; a dashed arrow indicates no statistically significant effect.

from an average of 45% to 19%, a 26% difference; lower-performers' dropped from 24% to 11%, a 13% difference.

5.6.2 Student performance, engagement time, and the ability to assess answers correctly. We found no statistically significant interaction effects related to engagement time. For autograder error detection, we examined three aspects: (1) how class performance explained the overall ability to assess one's answer accurately, (2) whether encountering an autograder error impaired higher and lower-performing participants' judgments equally, and (3) whether error detection had differential downstream effects on how likely one learned depending on one's performance.

For the first aspect, lower-performers were more likely to judge correctly whether their answers were right or wrong with a small effect size (Odds ratio comparing higher vs. lower-performers; FP: M = 0.68, 94% HPDI = [0.49, 0.88]; FN: M = 0.62, 94% HPDI = [0.39, 0.84]). This translates to an absolute difference in probabilities of about 5-10% – for example, lower-performing survey participants detected FPs with an average chance of 28%, vs. 18% for higher-performers.

For the second aspect, we found no significant interaction effects. For instance, comparing TNs to FPs given one wrote a wrong answer, FPs lowered the odds of recognizing one's mistakes by about 50% for everybody.

Lastly, we found that the detection of autograder errors benefited higher-performers' learning more than lower-performers, with a small effect size. In these conditions, higher-performers on average had a 43% higher chance to answer a post-test question correctly (Odds ratio comparing higher vs. lower performers: 94% HPDI = [1.00, 1.95]).

5.6.3 A special case of a low performer. While we made no systematic comparisons of behavior or themes between interviewees of different self-reported grades, one case stood out. P4 demonstrated a much lower understanding of Python code than the other participants and wrote only incorrect answers, both on the practice questions and post-test. They said that they gained little from the feedback, even when acknowledging their answer was wrong: *I read the answers and they're not what I put, so yeah. I guess I understand it a little bit, but not really.* Later, their compared reading the autograder's feedback to listening to a lecture: *If I were to learn this stuff, I have a better time with my tutor one-on-one ... 'Cause that way I can talk it through with somebody who actually knows what they're doing, instead of just listening to [a lecture].*

P4 admitted they did not have the knowledge to tell whether the grader was correct. This caused them to agree with the autograder by default, including during FPs. But even if they had assessed their answer correctly, P4's general interaction with the autograder's feedback suggests they still would be unlikely to learn.

6 DISCUSSION AND FUTURE WORK

Based on our findings, we propose suggestions to improve the design and workflow of the ASAG used in our experiment, as well as other ASAGs that rely on a similar style of feedback. Additionally, we use our findings to motivate future work in AI feedback systems more broadly.

6.1 Managing Wrong Answers Marked as Correct (FPs)

We found two primary drivers that decreased the likelihood that participants detected FPs, resulting in learning harms. The grader's positive mark caused 1) participants to not pay attention to feedback, and 2) a bias against admitting mistakes in their answers. One proposal to mitigate the first driver is to increase attention to FPs, such as by warning students that FPs exist, as prior work found many of them were unaware [33], or providing incentives like extra credit to those that identify FPs. Yet, examining a positive mark acts against students' instincts, and not everybody responds to such incentives. Moreover, we wish to avoid students not gaining anything after examining feedback on contents they already know well.

Instead of incentivizing engagement with positive feedback all the time, we advocate for a more selective approach. We suggest prioritizing higher-performing students, as we found that they detected FPs less compared to lower-performers on the survey. We further hypothesize that higher-performers were more confident in their answers and therefore believed the autograder more when marked as correct. If this is true, then we recommend incentivizing engagement for students that display unwarranted confidence. Future work can explore other groups of students and test the efficacy of encouraging engagement with positive feedback in mitigating harm from FPs.

Another suggestion, which may also help overcome bias against admitting one's own mistakes, is to increase detail in positive feedback. Detailed positive feedback in our context could explain how a response fulfills class standards and expresses important concepts, such as *Check successful: the autograder thinks your answer unambiguously specifies the inputs when it says "list x."* This will increase the visibility of contradictions caused by autograder errors and directly addresses participants' habits of shallow comparison with the exemplar answers. Furthermore, detailed positive feedback can help students confirm their knowledge even when their answers are truly correct, improving learning [51].

6.2 Managing Right Answers Marked as Incorrect (FNs)

Unlike interview participants, survey participants were less likely to answer post-test questions correctly after a FN. This effect was explained neither by variations in engagement time within each group of participants, nor by participants' failures to assess their answers as correct. With those explanations ruled out, we propose that the guided reflection process unique to the interviewees protected them from learning harms. Interviewees had to describe their use of the feedback and articulate why they thought their answer was right or wrong. Such back-and-forth engagement, with many elements of the *reflect when prompted* technique, could have increased students' cognitive engagement and encouraged extra metacognition. While the benefits of these processes for learning outcomes are well-known [7], our findings now suggest the additional tangible benefit of reducing harm from grading errors.

An implementation of the *reflect when prompted* technique could ask students who are confused with an autograder's wrong mark to externalize their reflection on why they think their answer is correct and how it meets the class standards by completing a checklist of criteria that correct responses must fulfill [55]. Another approach involves rethinking workflows around the autograder – designing an online forum to discuss AI output with others (e.g., peers), similar to Duolingo's discussion pages for machine-graded translation tasks. We leave the implementation and comparison of various strategies to future work.

6.3 Deciding FN vs. FPs Trade-off in AI Short Answer Autograding

Classifiers inherently make trade-offs; decreasing the FP rate usually increases the FN rate, and vice-versa. Common sense dictates that the ratio of FPs versus FNs should be guided by the goal of the AI application and the consequences of errors. Here we describe how educators may use our findings to inform their decisions.

First, we observe that after experiencing FNs, no interviewees expressed incorrect coding concepts, and participants' post-test responses were often incorrect because of unsuccessful rewording attempts, such as attempts that used less-specific language. In contrast, FPs prevented participants from correcting their conceptual errors in many cases. In other words, FPs harmed conceptual understanding more than FNs.

Second, our suggestions for mitigating FNs already match best practices for engaging with feedback. But mitigating FPs seems more challenging because of the need to work against students' tendencies to disengage with feedback and their bias against admitting their errors after receiving a positive mark. Additionally, encouraging engagement with FPs may risk having students gain nothing from studying content they already know well with TPs.

And finally, prior work on a similar AI autograder suggested that students look upon FNs in low-stakes assessments more leniently than FNs in high-stakes assessments [33]. Thus, as long as AI is deployed in low-stakes formative assessments, instructors have some leeway in biasing towards a higher ratio of FNs while keeping students reasonably satisfied.

Considering all of the above, when mastering course materials is the primary goal, such as our study's formative setting within a for-credit course in secondary education, we suggest prioritizing the reduction in FP rate. At the same time, we recommend leveraging common mechanisms of formative assessments [10, 15, 68] to reduce dissatisfaction with FNs [33]. Notably, without evidence from our study, the instructors of the course in our study chose the opposite – to reduce FPs, since students' dissatisfaction with FNs was more apparent and well-understood than the harms from FPs. Taken as a case study, our study shows the limitations of intuition and the need for empirical evaluation of the impact of FPs and FNs. In other contexts with different stakes and priorities, such as when boosting learners' curiosity and motivation takes priority over mastery, we encourage educators to take a similar approach to the one we have taken to make FP-FN trade-off decisions.

6.4 Designing for Different Student Populations

We found no evidence that autograder errors harmed lower-performing students more than higher-performing students, contrary to our original hypothesis. In fact, higher-performers had a bigger absolute

decrease in learning outcomes after experiencing a FP in the survey condition. We acknowledge the possibility that lower-performers' learning outcomes were already low enough to limit the magnitude of further drops. However, higher-performing participants still had more trouble accurately assessing their answers' correctness, which caused lower learning outcomes for FPs. Thus we suggested targeted interventions to help higher-performers detect FPs in Section 6.1.

Lower-performers, on the other hand, overall assessed their answers' correctness more accurately. But when both lower- and higher-performers accurately assessed their answers, lower-performers were less likely to learn. As the example in Section 5.6.3 suggests, detecting an autograder error does not matter if the student cannot learn from the provided feedback. In other words, lower-performers do not need as much help detecting autograder errors, but could benefit more from our before-mentioned suggestions for detailed feedback (Section 6.1) and enhancing cognitive engagement (Section 6.2).

Besides performance, prior work shows that students with different self-efficacy, motivational beliefs [63], gender [64], etc., interact with and learn from feedback differently. Hence, we encourage future research to study how autograder errors impact these populations differently and consider these variations when designing AI autograders to enhance fairness.

6.5 What About More Sophisticated AI Autograders?

The ASAG in our study used a simple bag-of-words model. But future ASAGs will likely be more sophisticated. Currently, we are seeing much commentary on large language models (LLMs), such as OpenAI's ChatGPT [50] and Google's BARD [27]. LLM technology offers promise for autograders to provide human-like interactions and detailed feedback. However, current designs of LLMs provide no signals of which outputs could be false – people still need to decide when and how much to trust the AI's feedback. The fluency of these responses may lead users to think that a human generated the responses [35], masking their inaccuracies. Furthermore, LLM's false feedback may not cleanly fit into the categories of FPs and FNs. Given the differential learning impacts of FPs and FNs in our study, we call for future research to create a taxonomy for the errors in LLM-generated feedback, examine how well users detect these errors, and evaluate their learning impacts.

6.6 Deploying AI Ethically

Finally, we must point out that while we focused on how AI errors affected learning in this paper, errors can harm other metrics and pose other serious consequences. For instance, errors or the potential of errors can prompt accusations of unfairness, such as when officials decided to use an algorithm that may have disproportionately benefited students from private schools to predict and determine students' A-level exam scores in the UK in 2020 [9]. In short, even when we can mitigate the learning harms of AI errors in contexts similar to our study, practitioners must still support consistency, fairness, and robust appeal processes to ethically and responsibly deploy AI in education.

7 LIMITATIONS

One limitation is that our dependent variables capture limited facets of how students engaged and learned. We chose to measure time spent on feedback due to its ease of measurement and objectiveness, but it does not include nuances such as *how* participants spent their time and how they engaged cognitively. We captured some of this in the interviews, but future work should look at other types of engagement more systematically. Likewise, our measure of "learning" captured if participants corrected their mistakes on the practice problems and whether they picked up new misconceptions. It did not necessarily evaluate whether they gained a deep understanding of code reading or writing. Future work should devise more sophisticated measures of learning, such as longer-term learning outcomes or skill transfer to code-writing ability and more.

Another limitation is that we derived our models of student behavior and subsequent recommendations from an introductory computer science class context and an autograding system used in this class. We do not claim that our exact estimates of causal effects will always generalize, especially in contexts with a significantly different feedback design or subjects where the obviousness of grading errors differs. Still, we believe most of our recommendations are relevant in a broad sense, especially around how to deal with FPs and how to determine FP-FN trade-offs, as we expect students to feel less incentive to engage with positive feedback in many contexts.

8 CONCLUSION

In this work, we investigated the impact of short-answer AI autograder errors on learning, the mechanisms of the impacts, and how the impacts differ according to student performance. We summarize our key messages below:

- Help learners detect FPs. Suggestions include encouraging engagement with positive feedback, especially for higher-performers, and increasing detail in all positive feedback.
- Enhance students' cognitive engagement and metacognition to mitigate both FNs and FPs. One approach is to have learners articulate answers' correct/incorrect aspects, either with a tool or other humans (e.g., peers).
- Personalize interventions based on student performance. Higher performers need more help with error detection, while lower performers need more guidance on learning from the feedback.
- Decrease the ratio of FPs to optimize for learning. This is because FNs harmed our participants' learning less than FPs, and current knowledge and best practices suggest easier mitigation for FNs.
- Empirically evaluate the effects of FPs and FNs. We demonstrate the effectiveness of this approach in this paper, and encourage similar evaluations when deploying AI educational systems in new contexts.

Looking forward, we recommend investigating the impact of erroneous AI-provided feedback in other contexts as these systems evolve and developing solutions to help learners get the most out of the feedback despite AI errors.

ACKNOWLEDGMENTS

A big thanks to the students that participated in our study. Additional thanks to the members of the Social Spaces group for their writing suggestions and feedback. This work was supported by the NSF (DUE 21-21424, NSF IIS-2016908), Capital One, and the UIUC College of Engineering Strategic Research Initiatives (SRI) grant. Tiffany Wenting Li was additionally supported by the Google Ph.D. fellowship.

REFERENCES

[1] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences* 4, 2 (1995), 167–207.

[2] Yigal Attali and Don Powers. 2008. Effect of immediate feedback and revision on psychometric properties of open-ended GRE® subject test items. *ETS Research Report Series* 2008, 1 (2008), i–23.

[3] Sushmita Azad, Binglin Chen, Maxwell Fowler, Matthew West, and Craig Zilles. 2020. Strategies for Deploying Unreliable AI Graders in High-Transparency High-Stakes Exams. In *International Conference on Artificial Intelligence in Education*. Springer, Springer International Publishing, Cham, 16–28.

[4] Albert Bandura. 1991. Social cognitive theory of self-regulation. *Organizational behavior and human decision processes* 50, 2 (1991), 248–287.

[5] Albert Bandura and Daniel Cervone. 1983. Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of personality and social psychology* 45, 5 (1983), 1017.

[6] Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of educational research* 61, 2 (1991), 213–238.

[7] Maria Bannert and Christoph Mengelkamp. 2008. Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning* 3, 1 (2008), 39–58.

[8] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.

[9] BBC. 2020. A-levels and GCSEs: How did the exam algorithm work? https://www.bbc.com/news/explainers-53807730. Accessed 2022-09-14.

[10] Randy Elliot Bennett. 2011. Formative assessment: A critical review. *Assessment in education: principles, policy & practice* 18, 1 (2011), 5–25.

[11] J Martin Bland and Douglas G Altman. 2000. The odds ratio. *Bmj* 320, 7247 (2000), 1468.

[12] Denys Brand, Matthew D Novak, Florence D DiGennaro Reed, and Samara A Tortolero. 2020. Examining the effects of feedback accuracy and timing on skill acquisition. *Journal of organizational behavior management* 40, 1-2 (2020), 3–18.

[13] Joan F Brett and Leanne E Atwater. 2001. 360° feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied psychology* 86, 5 (2001), 930.

[14] Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education* 25, 1 (2015), 60–117.

[15] Jennifer C. Jacoby, Sheelagh Heugh, Christopher Bax, and Christopher Branford-White. 2014. Enhancing learning through formative assessment. *Innovations in Education and Teaching International* 51, 1 (2014), 72–83.

[16] Henian Chen, Patricia Cohen, and Sophie Chen. 2010. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—simulation and Computation®* 39, 4 (2010), 860–864.

[17] Roy B Clariana. 1990. A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction* (1990).

[18] Albert T Corbett and John R Anderson. 2001. Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 245–252.

[19] Malcolm Corney, Sue Fitzgerald, Brian Hanks, Raymond Lister, Renee McCauley, and Laurie Murphy. 2014. 'Explain in Plain English' Questions Revisited: Data Structures Problems. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* (Atlanta, Georgia, USA) *(SIGCSE '14)*. ACM, New York, NY, USA, 591–596. http://doi.acm.org/10.1145/2538862.2538911

[20] Donald B Fedor. 1991. Recipient responses to performance feedback: A proposed model and its implications. *Research in personnel and human resources management* 9, 73 (1991), 120.

[21] Lot Fonteyne, Annick Eelbode, Isabelle Lanszweert, Elisabeth Roels, Stijn Schelfhout, Wouter Duyck, and Filip De Fruyt. 2018. Career goal engagement following negative feedback: Influence of expectancy-value and perceived feedback accuracy. *International journal for educational and vocational guidance* 18, 2 (2018), 165–180.

[22] Max Fowler, Binglin Chen, Sushmita Azad, Matthew West, and Craig Zilles. 2021. Autograding" Explain in Plain English" questions using NLP. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1163–1169.

[23] Patricia Gaynor. 1981. The effect of feedback delay on retention of computer-based mathematical material. *Journal of Computer-Based Instruction* 8, 2 (1981), 28–34.

[24] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

[25] Andrew Gelman and Donald B Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical science* (1992), 457–472.

[26] Graham Gibbs and Claire Simpson. 2005. Conditions under which assessment supports students' learning. *Learning and teaching in higher education* 1 (2005), 3–31.

[27] Nico Grant and Cade Metz. 2023. Google Releases Bard, Its Competitor in the Race to Create A.I. Chatbots. https://www.nytimes.com/2023/03/21/technology/google-bard-chatbot.html. Accessed 2023-3-24.

[28] Douglas Grimes and Mark Warschauer. 2010. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment* 8, 6 (2010), 44 pages.

[29] Karen Handley, Margaret Price, and Jill Millar. 2011. Beyond 'doing time': investigating the concept of student engagement with feedback. *Oxford Review of Education* 37, 4 (2011), 543–560.

[30] Gerald S Hanna. 1976. Effects of total and partial feedback in multiple-choice testing upon learning. *The Journal of Educational Research* 69, 5 (1976), 202–205.

[31] Richard Higgins, Peter Hartley, and Alan Skelton. 2002. The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in higher education* 27, 1 (2002), 53–64.

[32] Jason M Hirst and Florence D DiGennaro Reed. 2015. An examination of the effects of feedback accuracy on academic task acquisition in analogue settings. *The Psychological Record* 65, 1 (2015), 49–65.

[33] Silas Hsu, Tiffany Wenting Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. 2021. Attitudes Surrounding an Imperfect AI Autograder. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 681, 15 pages. https://doi.org/10.1145/3411764.3445424

[34] Daniel R Ilgen, Cynthia D Fisher, and M Susan Taylor. 1979. Consequences of individual feedback on behavior in organizations. *Journal of applied psychology* 64, 4 (1979), 349.

[35] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120.

[36] Douglas A Johnson, Jessica M Rocheleau, and Rachael E Tilka. 2015. Considerations in feedback delivery: The role of accuracy and type of evaluation. *Journal of Organizational Behavior Management* 35, 3-4 (2015), 240–258.

[37] Anders Jonsson. 2013. Facilitating productive use of feedback in higher education. *Active learning in higher education* 14, 1 (2013), 63–76.

[38] Sally Jordan. 2012. Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions. *Computers & Education* 58, 2 (2012), 818–834. https://doi.org/10.1016/j.compedu.2011.10.007

[39] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4521–4532.

[40] Angelo J Kinicki, Gregory E Prussia, Bin Joshua Wu, and Frances M McKee-Ryan. 2004. A covariance structure analysis of employees' response to performance feedback. *Journal of applied psychology* 89, 6 (2004), 1057.

[41] Christoph König and Rens van de Schoot. 2018. Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review* 70, 4 (2018), 486–509.

[42] Raymond W Kulhavy, Mary T White, Bruce W Topp, Ann L Chan, and James Adams. 1985. Feedback complexity and corrective efficiency. *Contemporary educational psychology* 10, 3 (1985), 285–291.

[43] Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities* 37, 4 (2003), 389–405.

[44] Raymond Lister, Colin Fidge, and Donna Teague. 2009. Further Evidence of a Relationship Between Explaining, Tracing and Writing Skills in Introductory Programming. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education* (Paris, France) *(ITiCSE '09)*. ACM, New York, NY, USA, 161–165. https://doi.org/10.1145/1562877.1562930

[45] Mike Lopez, Jacqueline Whalley, Phil Robbins, and Raymond Lister. 2008. Relationships between reading, tracing and writing skills in introductory programming. In *Proceedings of the Fourth International Workshop on Computing Education Research*. ACM, 101–112.

[46] Uwe Maier, Nicole Wolf, and Christoph Randler. 2016. Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic

items and different feedback types. *Computers & Education* 95 (2016), 85–98.
[47] B Jean Mason and Roger Bruning. 2001. Providing feedback in computer-based instruction: What the research tells us. *Retrieved February* 15 (2001), 2007.
[48] Santosh A Mathan and Kenneth R Koedinger. 2002. An empirical assessment of comprehension fostering features in an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems*. Springer, 330–343.
[49] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
[50] Cade Metz. 2022. The New Chatbots Could Change the World. Can You Trust Them? https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html. Accessed 2022-12-14.
[51] Antonija Mitrovic, Stellan Ohlsson, and Devon K Barrow. 2013. The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers & Education* 60, 1 (2013), 264–272.
[52] Edna H Mory. 1994. Adaptive feedback in computer-based instruction: Effects of response certitude on performance, feedback-study time, and efficiency. *Journal of Educational Computing Research* 11, 3 (1994), 263–290.
[53] Laurie Murphy, Renée McCauley, and Sue Fitzgerald. 2012. 'Explain in Plain English' Questions: Implications for Teaching. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education* (Raleigh, North Carolina, USA) *(SIGCSE '12)*. ACM, New York, NY, USA, 385–390. https://doi.org/10.1145/2157136.2157249
[54] Susanne Narciss, Sergey Sosnovsky, Lenka Schnaubert, Eric Andrès, Anja Eichelmann, George Goguadze, and Erica Melis. 2014. Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education* 71 (2014), 56–76.
[55] David Nicol. 2021. The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education* 46, 5 (2021), 756–778.
[56] Stephen R Porter, Michael E Whitcomb, and William H Weitzer. 2004. Multiple surveys of students and survey fatigue. *New directions for institutional research* 2004, 121 (2004), 63–73.
[57] Doris R Pridemore and James D Klein. 1995. Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology* 20, 4 (1995), 444–450.
[58] Barry J Schimmel. 1983. A Meta-Analysis of Feedback to Learners in Computerized and Programmed Instruction. (1983).
[59] Marvin L Schroth. 1992. The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary educational psychology* 17, 1 (1992), 78–82.
[60] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.
[61] Hazel K Sinclair and Jennifer A Cleland. 2007. Undergraduate medical students: who seeks formative feedback? *Medical education* 41, 6 (2007), 580–582.
[62] D Sleeman, Anthony E. Kelly, R Martinak, Robert D Ward, and Joi L Moore. 1989. Studies of diagnosis and remediation with high school algebra students. *Cognitive science* 13, 4 (1989), 551–568.
[63] Caroline F Timmers, Jannie Braber-Van Den Broek, and StéPhanie M Van Den Berg. 2013. Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & education* 60, 1 (2013), 25–31.
[64] Gill Turner and Graham Gibbs. 2010. Are assessment environments gendered? An analysis of the learning responses of male and female students to different assessment environments. *Assessment & Evaluation in Higher Education* 35, 6 (2010), 687–698.
[65] Don VandeWalle and Larry L Cummings. 1997. A test of the influence of goal orientation on the feedback-seeking process. *Journal of applied psychology* 82, 3 (1997), 390.
[66] Anne Venables, Grace Tan, and Raymond Lister. 2009. A Closer Look at Tracing, Explaining and Code Writing Skills in the Novice Programmer. In *Proceedings of the Fifth International workshop on Computing Education Research*. ACM, 117–128.
[67] Jacqueline Whalley, Raymond Lister, Errol Thompson, Tony Clear, Phil Robbins, P K Ajith Kumar, and Christine Prasad. 2006. An Australasian study of Reading and Comprehension Skills in Novice Programmers, using the Bloom and SOLO Taxonomies. *Eighth Australasian Computing Education Conference (ACE2006)* (2006).
[68] Dylan Wiliam. 2007. Keeping learning on track: Classroom assessment and the regulation of learning. Information Age Publishing.
[69] Sue Ellen Williams. 1997. Teachers' written comments and students' responses: A socially constructed interaction. (1997).
[70] Naomi E. Winstone, Robert A. Nash, Michael Parker, and James Rowntree. 2017. Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes. *Educational Psychologist* 52, 1 (2017), 17–37. https://doi.org/10.1080/00461520.2016.1207538 arXiv:https://doi.org/10.1080/00461520.2016.1207538

9 APPENDIX

9.1 Group 2 Confounders

These factors could reasonably and simultaneously predict engagement with feedback, beliefs about the correctness of one's response, and/or learning outcomes. We present reasonings and evidence that each confound affects at least two of these outcomes.

(1) Task-value belief. Prior work found that the perceived value of a task predicted the effort participants invested in a formative assessment about that task and feedback-seeking behavior [63].

(2) Task success expectancy. People's perceptions of their likelihood to do well on a task predicts feedback-seeking behavior [63].

(3) Goal orientation. Prior work found that learning goal orientation predicts feedback-seeking behaviors, and a performance-avoidance goal orientation negatively predicts feedback-seeking behavior [65].

(5) Gender. Empirical evidence suggests that gender affects feedback usage and subsequent learning [46, 54].

(6) English proficiency. We reasoned that people less proficient in English would take longer to read the feedback. At the same time, we reasoned people less proficient in English may have more difficulty in constructing answers to EiPE problems. (Notably, participants in [33] had the same concern.)

(7) Prior trust of the autograder. This intuitively affects beliefs about the correctness of one's response after seeing autograder feedback, and we also reasoned that participants could ignore or skip feedback if they did not trust it.

(8a and 8b) In-experiment autograder behavior. Perceptions and trust of the autograder could change as students see the autograder's output in the experiment. We model autograder behavior with (A) the number of questions that have been marked as correct up to the current question and (B) the number of mistakes the autograder has made so far.