Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality

Xu Wang
University of Michigan
Ann Arbor, United States
xwanghci@umich.edu

Carolyn P. Rose
Carnegie Mellon University
Pittsburgh, United States
cprose@cs.cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University
Pittsburgh, United States
koedinger@cmu.edu

ABSTRACT

Maximizing system scalability and quality are sometimes at odds. This work provides an example showing scalability and quality can be achieved at the same time in instructional design, contrary to what instructors may believe or expect. We situate our study in the education of HCI methods, and provide suggestions to improve active learning within the HCI education community. While designing learning and assessment activities, many instructors face the choice of using open-ended or close-ended activities. Close-ended activities such as multiple-choice questions (MCQs) enable automated feedback to students. However, a survey with 22 HCI professors revealed a belief that MCQs are less valuable than open-ended questions, and thus, using them entails making a quality sacrifice in order to achieve scalability. A study with 178 students produced no evidence to support the teacher belief. This paper indicates more promise than concern in using MCQs for scalable instruction and assessment in at least some HCI domains.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**.

KEYWORDS

HCI education; instructor belief; learning@scale;learning experience design; multiple-choice questions; matched assessment comparison

ACM Reference Format:
Xu Wang, Carolyn P. Rose, and Kenneth R. Koedinger. 2021. Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama,Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445045

1 INTRODUCTION

Increasing numbers of people are seeking higher education through online and physical courses and programs. Solutions to meet this growing demand, e.g., learning management systems and Massive Open Online Courses, have placed substantial emphasis on technology solutions that are easy to scale. However, the scalable learning solutions that have been employed have come with the perception of lower quality, reinforcing the idea that the two goals of scalability and quality are at odds.

For example, online distribution of videotaped lectures is a powerful technique for scaling education but alone it is in conflict with research suggesting that more interactive forms of learning-by-doing produce higher quality learning [14, 24, 36]. As another possible example of this scale-quality trade-off, consider alternative ways to provide active learning opportunities online. Do assignments implemented via multiple-choice questions (MCQs) *provide for scale* because grading and instructional feedback can be easily automated *but sacrifice quality* relative to open-ended assignments where solution generation and human-generated feedback enhance the learning experience? Or might carefully designed MCQs provide rich learning experiences *and* offer the advantage of immediate feedback? In this paper, we address this tension between scalability and quality from both instructors' and students' perspectives and provide evidence that though some paths towards scalability have resulted in reductions in quality, it does not have to be that way.

MCQs can be graded automatically within many popular online learning and testing platforms, e.g., Canvas [20], GradeScope [43]. The benefit of MCQs also extends to Massive Open Online Courses, where grading and offering feedback to hundreds or thousands of students has been a substantial problem [19, 21, 27, 40]. One might argue, however, that though MCQs have practical value in terms of ease of grading, using them comes at a cost in terms of quality of insight provided. One temptingly sensible argument is "since recognition is easier than recall, MCQs are easier than open-ended questions thus do not exercise the same level of thinking." Another we have heard from instructors is "Open-ended questions exercise students' critical thinking skills while MCQs don't." In this paper, we provide both evidence for the prevalence of such beliefs and evidence for questioning these beliefs as they mismatch student performance data. We also present alternatives to the arguments above that provide theoretical reasons for why and when MCQs can provide for equivalent or better learning *quality* while enhancing prospects for *large-scale* support for learning by doing.

We situate our study in HCI education, with the goal of improving HCI pedagogy at scale. The rapid growth of the UX profession has led to an increased need for qualified practitioners and a proliferation of UX educational programs [37]. Recently, HCI educators have begun to reflect more on pedagogy and practice [55]. Prior work on HCI pedagogy is often designed to give students exposure to the entire lifespan of a UX project, e.g., through studio-based and project-based learning approaches [18, 26, 34, 37, 51]. These approaches were valuable in providing students with hands-on experiences and opportunities to interact with real users [37]. However,

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445045

from our own experiences as HCI educators and prior work[34], some HCI design and research projects can be overly challenging for novice students when they are not fluent with HCI skills yet and need to manage the extra cognitive efforts introduced by interacting with real users and project management. As an example, when students have not mastered how to write interview questions, asking them to conduct interviews with high-stake users may not be the best use of their and the participants' time. In this work, We investigate what are the most efficient ways to equip students with the necessary concepts and skills in HCI research and evaluation, as a first step before engaging students in project or studio-based learning.

We first surveyed HCI instructors to understand their beliefs and considerations when using multiple-choice and open-ended activities in their teaching. Prior work has found that teachers' interpretations and implementations of curricula (e.g., math) are greatly influenced by their knowledge and beliefs about instruction and student learning [8, 9, 31]. Thus it is important to examine the accuracy of instructor beliefs in response to students' actual performance. In a survey with 22 professors from 9 institutions that are teaching HCI research methods courses, participants showed a preference of using open-ended questions in their courses. The surveyed instructors tend to believe that MCQs are less valuable because recognition is easier than recall, and that open-ended questions exercise critical thinking whereas MCQs do not.

We next conducted study 2 in which we compared student performance data with instructor predictions regarding the difficulty of matched multiple-choice and open-ended questions. Student performance data is surprisingly at odds with instructors' beliefs. We designed 18 pairs of matched multiple-choice and open-ended questions on HCI research methods, including the topics of interview question design and think-aloud protocols, which are rated in a recent study [11] by HCI educators and practitioners as "very important" HCI design and empirical methods. A total of 178 students in two college courses answered these questions as a part of their exams. Student performance data contradicted the instructors' predictions. We found no evidence that open-ended questions were harder as predicted by instructors. At the same time we found substantial evidence that MCQs were not easy. The result supports the hypothesis that in the areas that we investigated, well-designed MCQs are assessing, and exercising during practice, the same difficult skills that are exercised in open-ended questions.

We suggest three general contributions of this work. First, our work indicates that, at least for some domains in HCI, online learning can benefit from the scaling advantages of multiple-choice questions without sacrificing (and perhaps gaining) learning quality. Learning experience (LX) designers may consider, with less guilt, the use of multiple-choice assessment and practice. To determine what subject-matter may have the required characteristics (e.g., evaluative skill is distinctly challenging), LX designers may use our matched assessment comparison technique to identify when MCQs are equally difficult.

Second, our work provides further evidence that instructors have so-called "expert blind spots", revealed through cases where their beliefs and student performance do not match [32, 33]. Instructor beliefs are important because they will influence the design of curriculum and learning experience of students. In both this and a past case [25], we see experts have good reasons for their beliefs, yet data suggests otherwise and a deeper analysis explains why. The instructor reasoning provided and the actual reasoning suggested by student performance data for both cases are displayed in Table 1. More generally, our work suggests that reasoning behind educational decisions can be probed through well-designed, low-effort, experimental comparisons toward more nuanced and accurate reasoning and decision making, and ultimately better design.

Third, our work surfaces a missing knowledge piece in instructional design especially in higher-education. College instructors are experts in their domains, but they are not necessarily experts on pedagogy. In many other domains, design of products to support the workflow of professionals require expertise from both domain experts and interaction designers, e.g., interaction designers design products to support doctors' decision making [57]. However, instructors are frequently required to take on both roles though their expertise does not prepare them for both. Our work suggests that, consistent with other design practices, to improve quality of learning design in higher-education, establishing roles such as learning designers or learning engineers is desirable.

2 RELATED WORK

In this section, we discuss the debate in prior work about the pros and cons of MCQs and open-ended questions for assessments and practice. Our work contributes to this literature about the potential use and design of MCQs in novel HCI content domains. We discuss prior work that aimed at understanding instructor beliefs in correlation with their instructional actions, which motivated our design of the studies to investigate whether instructor beliefs align with student performance and probe into the reasoning behind instructors' beliefs. We then discuss prior studies that used matched pairs of questions of different formats to investigate the relative difficulty between them. The methods used in prior work inspired the design and implementation of our study. We review prior work on HCI education and pedagogy and we discuss how our work complements to existing literature in equipping beginner HCI students with HCI research and evaluation skills. We finally review recent technology advances in learning at scale and suggest how the insights we have gained from this study could be integrated with existing technologies, inform the design of new technologies, and facilitate learning at scale in practice.

2.1 Debate Around the Use of Open-ended vs Multiple Choice Questions

Prior work has discussed the use of multiple-choice versus open-ended items in assessments, especially in STEM domains. There has been a debate around whether performance tasks can be cognitively authentic without being strictly hands-on. It is generally assumed that more "authentic" and costly methods of assessment, such as hands-on performance tasks in science, yield more valid estimates of student knowledge than do more efficient methods, such as paper-and-pencil multiple-choice items, although a number of authors (e.g., [38, 42]) suggest that certain assessment and practice activities can be cognitively authentic – that is, can elicit the kinds of cognitive processing characteristic of expertise in a domain – without being contextually authentic [47].

	Story problems vs. Equations (Koedinger & Nathan, 2004)	**MCQs vs. Open-ended questions (this work)**
Instructor beliefs	Story problems are harder than matched equations	Open-ended questions are harder than matched MCQs
Instructor reasoning	Because equations are needed to solve the story problem	Because recognition is easier than recall
Student data suggests	Equations are harder than matched story problems	MCQs are of similar difficulty as open-ended questions
Deeper analysis explains why	Story problems can be solved without equations and equations are harder to learn to read than appreciated	The distinctly hard skills that must be learned are evaluative skills required by both multiple-choice and open-ended questions and not the generative skills uniquely demanded by open-ended questions

Table 1: Two example cases where instructor beliefs and student performance do not match because the expert reasoning does not align with the underlying cognitive processes of the students. A deeper analysis suggests what is going on with the students.

Prior studies indicate mixed findings in comparing the relative difficulty of multiple-choice and open-ended questions as assessment items. Funk and Dickson found that students performed better on multiple-choice questions compared to open-ended ones in a college psychology class [17]. Surgue et al. found similar results in 7th and 8th grade physics class [47].

However, other work found competing results showing multiple-choice questions can be equally effective for learning compared to their open-ended counterparts, and even offer some advantages. For example, Smith and Karpicke found that students performed equally well on English reading tasks no matter whether they practiced with multiple-choice, short-answer, or hybrid questions [44]. Similarly, Little et al. found that multiple-choice questions provide a win-win situation compared to open-ended cued-recall tests on English reading tasks [29, 30]. The authors found that both open-ended and cued-recall tests foster retention of previously tested information, but multiple-choice tests also facilitated recall of information pertaining to incorrect alternatives, whereas cued-recall tests did not.

Beyond the reality that the debate around the use of MCQs and open-ended questions has not yet reached consensus, we also see that the studies discussed above have focused on learning objectives that fall into only a subset of categories of learning activities in Bloom's Taxonomy [5]. In particular, the categories of "Knowledge" and "Comprehension" (e.g., learn psychology concepts, comprehend English paragraphs) in Bloom's taxonomy have been explored, but questions remain about the remaining categories. Some of the tasks may touch upon "Application" (e.g., apply knowledge about voltage and resistance to solve the current in a circuit). Few studies have explored the merits and drawbacks of MCQs and open-ended questions for assessing and practicing learning objectives that involve "Analysis", "Synthesis" and "Evaluation." In this paper, we broaden the empirical foundation available to ground instructional design by investigating learning objectives that involve "Evaluation" of candidate solutions.

2.2 Importance of Instructor Belief

Prior work has found that teachers' interpretations and implementations of curricula (e.g., math) are greatly influenced by their knowledge and beliefs about instruction and student learning [31]. Thus it is important to examine the accuracy of instructor beliefs in response to students' actual performance. For example, Nathan and Koedinger asked high school teachers to rank order the relative difficulty of six types of mathematics problems and found that teachers accurately judged students' performance abilities on some types of problems but systematically misjudged them on others [32]. In another case, Brown and Altadmri [8, 9] found that educators were not good at estimating common student mistakes when learning Java programming. In our investigation of the use of MCQs and open-ended questions, we performed a survey with university instructors to understand their beliefs and specific judgments about the difficulty of matched pairs of multiple-choice and open-ended questions. This is the first work we know of that investigates university instructor beliefs on the use of MCQs versus open-ended questions and compares instructor judgments with student performance.

2.3 Relative Difficulty of Matched Questions

Prior work used matched pairs of assessment questions to investigate the relative difficulty of questions of different formats, which can shed light on whether one format of questions is more valuable for practice and assessment compared to others. For example, Surgue et al. compared the difference between a real hands-on task (e.g., assembling an electric circle) and a written analogue of the task [47]. The study found that mean scores on the hands-on and written analogue tests were very similar, suggesting written analogue tests can be interchangeable as hands-on tasks that require actual manipulation of equipment. Noreen Webb et al. did a similar comparison between matched hands-on and paper-and-pencil tasks and showed consistent results [54]. Koedinger and Nathan designed matched algebraic problems in three formats, story problem, word equation and symbolic equation [25]. They found that symbolic equations are harder than matched story problems and word problems. In our study, we adopted a similar approach to compare the relative difficulty of matched pairs of MCQs and open-ended questions.

2.4 HCI Pedagogy

Recently, researchers and educators around the world have been investing in efforts to cultivate an HCI education community, develop effective HCI curricula, and reflect on and improve HCI pedagogy

and techniques [1, 2, 6, 10, 12, 16, 34, 37, 41, 45, 48, 50, 51]. Notable directions include identifying important components for instruction and practice [10, 11, 34], developing studio-based learning approaches [18, 26, 51], promoting reflexive practices and encouraging students to work with users [37], facilitating multi-disciplinary collaboration [2], implementing targeted pedagogical techniques such as flipped classrooms [12] and giving students video coursework [50]. Most prior work on studio-based and project-based learning approaches aim at providing students with hands-on experiences and opportunities to interact with real users [37]. Students liked the flexibility in these learning experiences and found them to be more experiential than traditional learning methods [37]. On the other hand, prior work has also pointed out difficulties in HCI education faced by students. For example, students need to navigate the complexity in these projects, and students who have less design background may encounter difficulties in getting started on design work, collaborating with others, and project management, etc. [34]

We consider our work to be complementary to existing work on HCI pedagogy with a focus on scalable beginner training. We answer questions such as what are ways to help novice students learn basic HCI research concepts and skills more efficiently, which serves as a preparation before students can meaningfully engage in full-stack UX projects.

2.5 Technologies to Support the Design and Use of Multiple-choice Questions

In this work, we perform the student-facing experiment within an exam context. However, the exam was used to test a specific hypothesis, not to illustrate a suggested context of use. The goal of the study is to infer the potential value of using multiple-choice and open-ended problems as formative learning activities. On the one hand, prior work has explored a variety of question generation techniques to produce high quality multiple-choice questions at scale. Leveraging these existing techniques, the authoring of multiple-choice questions can be made easier. For example, UpGrade sources student-written open-ended work [52], automatic question generation approaches [3, 28] leverage natural language processing techniques and existing knowledge ontology databases, Peerwise [13] and Concept Inventory [35] produce questions specific to programming education. On the other hand, with prior work in intelligent tutoring systems and conversational agents, we also see promises of providing adaptive and collaborative learning support using multiple-choice questions. For example, multiple-choice questions can be presented to students in an adaptive order leveraging adaptive problem selection techniques such as intelligent tutoring systems [22, 39]. Multiple-choice questions may also be integrated into conversational agents and pedagogical agents that give students practice opportunities as they explore open-ended tasks and as they collaborate in groups [49, 53, 56].

3 STUDY 1: INSTRUCTOR BELIEF SURVEY

In order to understand instructors' beliefs about using multiple-choice versus open-ended questions in their teaching. We conducted a survey with instructors who are teaching university level HCI research methods courses. The survey is composed of two parts, the first part asks about instructors' general beliefs on using multiple-choice or open-ended questions in their teaching. The second part asks participants to predict the relative difficulty of pairs of questions. In this section, we only present findings of the first part of the survey.

The questions in part 1 of the survey is shown in Figure 1. Before sending out the survey, we piloted it with two faculty who teach at a top-ranked professional HCI program. We drew on findings from these pilot tests to improve the clarity of the survey. The online survey was deployed using Qualtrics.

3.1 Participants

We obtained a list of university professors who are teaching or have taught HCI research methods from their websites. We added the ACs of the "Learning, Education and Families" subcommittee of CHI 2019 to the list. We sent 110 invitations in total, 22 participants completed the first part of the survey. All 22 responses met our inclusion criteria and were kept for analyses. Specifically, all participants confirmed that they had taught at least one course on this topic before, and rated their expertise in the topic area as expert or knowledgeable. The results below therefore represent perspectives from 22 HCI educators, who were affiliated with 9 institutions in the United States (n=20), Europe (n=1), and Asia (n=1). While this number is still relatively small, we feel that it is representative enough to provide initial insights, especially since we are already seeing similar themes in the open-ended responses to Q2 (suggesting saturation).

3.2 Survey Analysis

We analyzed survey questions with single-selection responses using descriptive statistics (Questions 1, 3, 4, shown in Figure 1). We examined open-ended responses (Question 2, shown in Figure 1) with inductive thematic analysis [7]. The three authors collaboratively analyzed the survey's open-ended responses and summarized the themes presented below.

3.3 Findings

3.3.1 Descriptive Statistics. All participants indicated that they are experts or knowledgeable in the content domain (HCI research methods) and have taught at least one course on a relevant topic before.

In response to Question 1a "I would pick open-ended rather than multiple-choice because multiple-choice questions and open-ended questions teach different skills.", 50% of the instructors answered "Always" and "Mostly", and 45% answered "Depends." In response to Question 1b "I would pick open-ended rather than multiple-choice because open-ended assignments are a way to develop critical thinking, which is not entirely possible via multiple-choice questions." , 73% of the instructors answered "Always" and "Mostly", 23% answered "Depends." In response to Question 3, 60% of the instructors thought students would gain more from doing open-ended practice, and 14% thought students would gain more from multiple-choice practice, the rest thought the two were the same or it depends on the topic. We see that instructors display a preference towards using open-ended questions and believe them to be more valuable in some ways. In response to Question 4, 45% of the instructors considered

1. Please judge the following statements, assuming you are designing an assignment and are considering whether to present a question in an open-ended format or in a multiple-choice format.
I would pick open-ended rather than multiple-choice because:

a) Multiple-choice questions and open-ended questions teach different skills	Always	Mostly	Depends	Rarely	Never
b) Open-ended assignments are a way to develop critical thinking, which is not entirely possible via multiple-choice questions	Always	Mostly	Depends	Rarely	Never

2. What are the skills exercised by open-ended problems and multiple-choice problems? Please describe your answer below.

3. For a given learning goal, which type of activities would allow students to gain more if they engage in the activity for the same amount of time? Compare students doing an hour of multiple-choice question practice and an hour of open-ended question practice.

Multiple-choie	Open-ended	Same	Others (Please explain)

4. For a given learning goal, which type of activities would be easier to design? Compare designing a multiple-choice practice problem and an open-ended practice problem

Multiple-choie	Open-ended	Same	Others (Please explain)

Figure 1: Survey questions that ask about instructors general belief about using MCQs and open-ended questions in their teaching.

open-ended questions to be easier to design, 23% considered MCQs to be easier to design, and the rest thought they were similar or it depends on situations.

3.3.2 Instructor Reasoning. In the survey, we asked instructors' views about the skills exercised by multiple-choice and open-ended questions respectively (Q2). Instructors mostly agree that multiple-choice questions test students' understanding of facts, whereas they gave diverse answers on the skills exercised by open-ended questions. Through a thematic analysis, we summarize three themes of answers as below.

Instructors tend to believe that MCQs mostly exercise recognition, and open-ended questions exercise recall.

> *"Multiple choice is mostly recognition over recall. They are also only good for questions that have a clear and well-defined answer. They also test knowledge, but not necessarily practice of skills." – P2*

Instructors think open-ended problems help students practice generating new ideas and development arguments, whereas multiple-choice questions help students test understanding of facts.

> *"Open-ended problems help students practice generating new ideas and developing arguments to support those ideas, which I see as key skills in HCI. Multiple-choice questions help students test their understanding of facts, and perhaps recognize good ideas or designs." – P12*

Instructors consider open-ended tasks to exercise critical thinking whereas multiple-choice do not.

> *"Open ended present better opportunities for students to exercise critical thinking and analytical thinking. It allows them to talk about relations and more abstract ideas (depending on the question). Multiple choice cannot do that. While they may encourage students to think, they mostly test students memory, possibly understanding, but rarely beyond that." – P18*

4 PROBING HYPOTHESES ABOUT THE BENEFITS OF OPEN-ENDED TASKS

In Study 1, we saw a preference of using open-ended tasks for teaching HCI-related concepts. Instructors consider open-ended tasks to be more valuable when teaching HCI-related concepts and they offer compelling reasons behind their choices. In this section, we derive alternative hypotheses about the benefits of open-ended tasks based on instructors' reasoning as displayed in Study 1 and prior work.

4.1 Hypothesis 1: Open-ended Tasks are Better Because They Exercise Extra Thinking Elements

As we observed in Study 1, instructors prefer to use open-ended tasks because they tend to think open-ended problems exercise critical thinking, idea generation and development and recall of information that multiple-choice questions do not exercise. These lines of reasoning suggest that open-ended tasks exercise extra thinking elements compared to multiple-choice tasks. The extra thinking elements could be recall of information, generation, or critical thinking, as shown in Figure 2

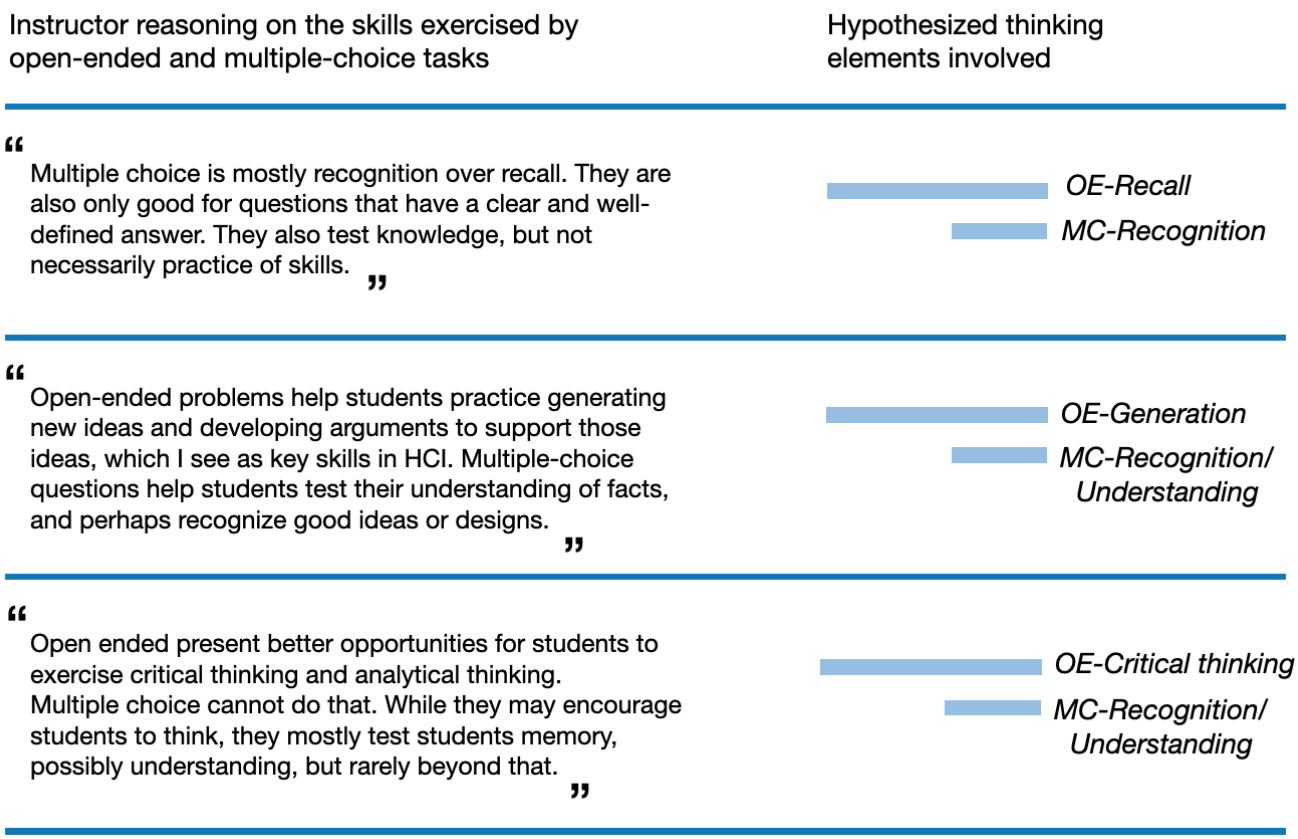

Figure 2: Instructors have different hypotheses about the thinking elements involved when answering multiple-choice and open-ended tasks, e.g., recognition vs. recall (Quote 1), recognition vs. critical thinking (Quote 3). All three cases suggest that instructors hypothesize there is a difference in the thinking elements involved when students attempt at these tasks, and they hypothesize open-ended tasks to involve more thinking elements.

4.2 Hypothesis 2: Open-ended and Multiple-choice Tasks Exercise Similar Thinking Elements

Prior work has shown that in some domains multiple-choice-type tasks can be as valuable for learning as open-ended tasks. For example, Yannier et al. [58] shows that evaluating "which towers would likely to fall" can be more effective in teaching kids physics principles around gravity and balance compared to having kids continuously build towers with LEGO . Wang et al. [52] shows that evaluating candidate solutions is equally effective in teaching college students how to design good survey questions compared to having students practice through generating survey questions. Ericson et al. [15] shows that when teaching programming, having students solve Parsons problems, i.e., evaluating the correctness and ordering of code snippets is equally effective for learning compared to having them write the equivalent code.

This line of work is motivating an alternative to Hypothesis 1 suggesting that multiple-choice tasks may be exercising similar thinking elements as open-ended tasks in some cases, which makes them equally beneficial for learning.

4.3 Using Difficulty of Matched Pairs of Questions to Test Hypotheses 1 and 2

We have derived two competing hypotheses around the relative thinking elements required in answering multiple-choice and open-ended questions. To test the hypotheses, we employed a method using difficulty of matched pairs of questions.

If Hypothesis 1 is true, it suggests that open-ended tasks exercise extra thinking elements that is eliminated in multiple-choice tasks. If the extra thinking elements are not mastered by novices, that will show up in open-ended tasks, but not in multiple-choice tasks. If we give students matched pairs of open-ended and multiple-choice tasks, the result for performance on these questions would be that the students get the multiple-choice right, even when they are missing the extra skills, but they will not get the open-ended right.

On the other hand, if Hypothesis 2 is true, which suggests that open-ended and multiple-choice tasks may exercise similar thinking elements. The result for performance on these questions would be that students get both versions of questions right or wrong at the same time, since the thinking elements required are similar.

Following this line of reasoning, giving students matched pairs of multiple-choice and open-ended questions and examining their performance on them will help us find out which hypothesis (1 or 2) is true.

5 STUDY 2: STUDENT PERFORMANCE ON MATCHED PAIRS OF QUESTIONS

We then conducted Study 2. There are two research goals in Study 2. First, we collect student responses on matched pairs of multiple-choice and open-ended questions to test the competing hypotheses above. Second, we compare student performance data with instructor prediction. Collecting student responses to matched pairs of questions helps us uncover the thinking elements required when answering these questions, which is evidence for instructors to make instructional decisions in their teaching. Checking the alignment between student responses and instructor prediction is also important, especially when they do not align, because instructors' beliefs greatly influence their implementations of curricula and student learning.

Study 2 was conducted within an exam context. However, the exam was used to test the above competing hypotheses, not to illustrate a suggested context of use. Collecting the data in a summative setting allowed us to better focus on the thinking elements of the two problem types and reduce confounding factors. e.g., students are less tempted to guess in a summative setting, and there is a lower attrition rate. With this understanding, we will further infer the potential value of using multiple-choice and open-ended questions as formative learning activities.

5.1 Study Context

We did Study 2 in two introductory HCI classes at an R1 institution in the United States. Both HCI classes cover a range of HCI empirical research methods. Based on the Churchil et al. (2016) [11] study, we selected 4 topics that are rated as "very important" and "important" components in HCI education, including conducting heuristic evaluation (usability inspection method that helps designers to identify usability problems in the user interface design and propose redesign features to address the problems), designing interview questions, interpreting notes from contextual inquiry interviews, and performing think-aloud studies.

5.2 Design of Matched Pairs of Questions

We designed 18 pairs of matched multiple-choice and open-ended questions in collaboration with the course instructors. Each pair of questions has the same question stem, the only difference is that the multiple-choice version offers options for students to choose from. The options in the multiple-choice version are designed using past students' mistakes. Example pairs of multiple-choice and open-ended questions used are shown in Figure 3. Multiple-choice questions use a different verb from the matched open-ended questions, e.g., suggest (for open-ended) versus select (for multiple-choice), and have 3 to 5 options for students to choose from, as shown in italics.

5.3 Study Design and Implementation

We performed an experiment in two separate classes to examine the relative difficulty between matched multiple-choice and open-ended questions. The two courses are both offered at a top-ranked professional Human-Computer Interaction (HCI) program in an R1 institution. Both courses cover HCI research methods, such as conducting interviews, and performing think-aloud protocols. We refer to the two courses as UX1 and UX2 for the rest of the paper.

Among the 18 pairs of questions, 4 pairs were used in UX1's mid-term exam, and 14 pairs were used in UX2's final exam. Taking UX1 as an example, with 4 pairs there are 8 question items in total. The 8 questions items are distributed into 2 exam forms. Form A contains Q1(MC)-Q2(OE)-Q3(OE)-Q4(MC), and Form B contains Q1(OE)-Q2(MC)-Q3(MC)-Q4(OE). In the design, we made sure Q1 and Q3 are testing the same knowledge component [23], while Q2 and Q4 are testing the same knowledge component. In this case, every student experienced both question formats for a given knowledge component. The two exams forms were randomly distributed among 103 students on exam day. For UX2, similar to UX1, two exam forms were created based on the 28 question items. We also made sure there were at least 2 questions on the same knowledge component, so that each student got to experience both question formats. The two exam forms were randomly distributed among 75 students on the exam day.

5.4 Answer Grading and Dataset

103 students from UX1 participated in the study. 49 of them did exam form A and 54 did exam form B. 75 students from UX2 participated in the study. 38 of them did exam form A and 37 did exam form B. Exams were graded as normal. One researcher and the course instructors collaboratively graded the exam answers. For multiple-choice questions, there is one correct answer, 1 being correct and 0 being incorrect. For open-ended questions, we used a strict grading criteria. The correct answer has to be the same or a close rephrase of the correct answer intended for the multiple-choice question, with 1 being correct and 0 being incorrect. 0.5 point were occasionally given (32 out of 1406 cases, 2%) to answers that addressed the intended problem but displayed additional errors. For example., consider a question asking students to revise an interview question that has the problem of asking secondhand information, if the student answer addressed this issue but displayed other errors, such as the new question is a leading question, this response is given 0.5 point. In UX1, 2 questions ask students to identify a heuristic violation of an interface, and the other 2 questions ask students to redesign the interface based on the problems. The answer of the latter question depends on their answer of the former question. To make the comparison fair, for the latter 2 questions, we only included students who answered the former question correctly in the dataset (regardless of question format).

For modeling and interpretation purposes, we removed the 32 entries with a score of 0.5. This results in a dataset of 1374 observations by 178 students. Each observation is a student response to a question. It has features including student ID, question ID, question format (multiple-choice or open-ended), and score (0 or 1). In the following section we present findings on this dataset. Here is a side note that, following this analysis, we did a second analysis treating

ID	Matched Pairs of MCQ and Open-ended Question	ID	Matched Pairs of MCQ and Open-ended Question
1	**Please suggest (select) one alternative question to improve the following interview question:** "What went wrong when you tried to open the file? Did it tell you it was corrupted?" *A. What went wrong when you tried to open the file? Did it tell you it can't be opened because it expired?* *B. Did anything go wrong when you tried to open the file?* *C. What happened when you tried to open the file?*	2	**Please indicate if the note is OK or describe(select) what is wrong with the interpretation note:** "P1-35 Got a 404 error" *A. The note should refer to the participant* *B. The note is OK* *C. Not enough context is provided, e.g., why did they get the error* *D. It contains jargon "404" error*
3	**Please describe (select) the one most salient issue for the following interview question:** "How do you think we should make this user interface more clear?" *A. No definition of "clear"* *B. Shouldn't ask users to suggest ideas* *C. Shouldn't ask about causality*	4	**Please describe (select) the one most salient issue for the following interview question:** "How often do you weigh yourself on a scale?" *A. Shouldn't ask a very personal question* *B. Shouldn't ask participants to estimate* *C. Shouldn't ask because of health care privacy laws*
5	**Apply heuristic evaluation. Examine the screenshot below. Assume that a student wants to register for a course and comes to this page.** *[screenshot]* **Please name (select) the most salient heuristic rule this violates?** *A. Recognition rather than recall* *B. Visibility of system status* *C. Aesthetic and minimalist design* *D. Flexibility and efficiency of use* *E. Match between system and the real world*	6	**Apply heuristic evaluation. Examine the screenshot below. Assume that a student wants to register for a course and comes to this page.** *[screenshot]* **Propose a (Which of the following) redesign feature to (would) fix the violation of "Recognition rather than recall"?** *A. Display the "Register for a course" panel in the middle of the page* *B. Allow users to select courses and sections from a dropdown menu* *C. Only keep the "Search for a course to register" link on this page* *D. Display all available courses on this page* *E. Display help and documentation (e.g., steps to register for a course) in the middle of the page*
7	**When conducting a think aloud, what would you do if a participant gets stuck on a task for 20 seconds?** *A. Tell the participant how to move forward* *B. Wait longer before intervening* *C. Ask the participant to describe why they get stuck* *D. End the task early*	8	**Please indicate if the note is OK or describe (select) what is wrong with the interpretation note:** "P1-5 The cafeteria is dirty, and the servers sometimes make mistakes on students' orders." *A. "Sometimes" is vague* *B. It contains two different concepts* *C. "Students" is unnecessary* *D. The note is OK*
9	**Please suggest (select) one alternative question to improve the following interview question:** "Which of these smartphone cases do you think your colleagues would like the most?" *A. Which of these 2 smartphone cases would your colleague X like better?* *B. Which of these smartphone cases do you like the most?* *C. Can you rank these 3 smartphone cases' popularity among your colleagues?*	10	**Please indicate if the note is OK or describe (select) what is wrong with the interpretation note:** "P1-8 He is male" *A. The note is OK* *B. "He" is unnecessary* *C. This should go in the profile* *D. Not enough context is provided*

Figure 3: 10 pairs of matched multiple-choice and open-ended questions that were used in both the instructor belief survey and in the subsequent classroom experiment. The multiple-choice format shows the options in italics whereas the open-ended format only shows the question stem.

all 0.5 point entries as 0 and kept all 1406 observations, which is a stricter grading method for open-ended problems. We saw similar results in the second analysis. We will make the anonymized dataset available through a public repository (consented by our participants in the IRB).

5.5 Findings: Multiple-choice Questions Do Not Avoid the Hard Part

We built a mixed-effect logistic regression model, with question score (0 or 1) as the dependent variable, and question format (multiple-choice or open-ended) as the fixed effect. Considering different students may have different abilities in the course, we included a random intercept for each student. Considering different questions may be of different difficulty and the relative difficulty between the two questions formats might differ for different questions, we included a random slope and a random intercept for each of the question in the model. We used the lme4 R package [4] to build the model, and the formula is shown below:

$$score = question\_form + (1|student\_id) + (1 + question\_form|question\_id) \quad (1)$$

We found that the fixed factor question format does not have an effect on the question score ($z = 0.352, p = 0.725$). The fixed effect coefficient has an estimate of mean of 0.077, and a 95% confidence interval of $[-0.352, 0.506]$. The random effects show that the student intercept parameter has a variance of 0.287, the question intercept parameter has a variance of 0.606, and the question slope parameter has a variance of 0.277. We take a further look at the random slope coefficient for each question to see whether question format impacts different questions differently.

For a given question j, and for one student i, the above formula looks like (2), where β is the fixed effect coefficient, β_j is the random

slope for question j, α is the fixed intercept, and α_j and α_i are random intercepts at question and student levels.

$$logit(score) = (\beta + \beta_j) * question\_form + \alpha + \alpha_j + \alpha_i \quad (2)$$

When inspecting the effect of question format for each individual question, we can check whether $\beta + \beta_j$ is in the 95% confidence interval of the fixed effect parameter β. If not, that would suggest the effect of question format for that question differs from zero. Among the 18 questions, 4 questions have $\beta + \beta_j$ that exceeds the confidence interval of $[-0.352, 0.506]$. The β_j for these questions are 0.482, 0.499, 0.612 and 0.708 respectively. All four questions show the trend that the open-ended format of this question received higher scores than the multiple-choice format on average. For the rest of the 14 questions, adding the random coefficient to the fixed effect coefficient does not make it different from zero, suggesting both formats of the questions are of similar difficulty. From this experiment, we do not observe a difference in the relative difficulty between matched pairs of multiple-choice and open-ended questions. In some cases, the trend shows that multiple-choice questions could be harder for students to answer compared to matched open-ended ones, for example Q1, 2, 3 as shown in Figure 3 and Table 2.

6 STUDY 2: DO STUDENT PERFORMANCE AND INSTRUCTOR PREDICTIONS ALIGN?

6.1 Instructor Prediction Survey

As we have mentioned earlier, the instructor survey we sent out in Study 1 is composed of two parts. In part 1, we included questions asking about instructors' general beliefs about using multiple-choice and open-ended questions. In part 2, the survey asks instructors to predict the relative difficulty of 10 pairs of multiple-choice and open-ended questions. The 10 pairs were randomly selected from the 18 pairs we used in the classroom experiments and we made sure the 10 pairs cover all 4 topics, namely heuristic evaluation, designing interview questions, interpreting notes from contextual inquiry interviews, and performing think-aloud studies.

The reason for not using the whole set of 18 pairs is mainly time constraint. When we piloted the survey with two HCI faculty, it takes more than 30 minutes to complete. Since many of the pairs of questions are about the same knowledge and skills (e.g., survey question design), we think the 10 pairs are representative of the whole set used in the classroom experiments. The final 10 pairs used in the survey is shown in Figure 3. For each pair, we display both questions and ask the instructor to predict which one would be harder. Each question reads "Which of the above two problems do you think is harder? (Hard here means you think the students are less likely to get it correct.)", followed by options of "a is harder than b", "b is harder than a", and "a and b are of the same difficulty."

At the end of part 2, we inserted an open-ended question asking participants to reflect on their process of rating the difficulty of the matched pairs. The question reads: "Reflecting on your decision making process for the above 10 questions. What criteria did you use to decide which problem is harder? Did you see yourself following a trend? E.g., selecting one type (multiple-choice or open-ended) over the other for most questions? Why? Were there exceptions to the trend? Why did you choose the other type in those cases?"

6.2 Participants

A subset of participants in Study 1 participated in part 2 of the survey. Among the 22 participants, 18 completed part 2. The 18 participants were affiliated with 9 institutions in the United States (n=16), Europe (n=1), and Asia (n=1). All participants rated themselves as "Exerpt" or "Knowledgeable" on the subject area and have taught at least 1 HCI research and evaluation methods courses before.

6.3 Findings: Instructor Prediction and Student Performance Data Do not Align

In response to the 10 questions on which of the two problems they think is harder, for 66% of the time, instructors answered that the open-ended question is harder than the multiple-choice question; 18% of the time, they answered that the multiple-choice question is harder than the open-ended question; For 16% of the time, they thought multiple-choice and the open-ended question are of the same difficulty.

We compared instructor prediction of the relative difficulty of matched multiple-choice and open-ended questions with student performance data for each of the 10 pairs. Table 2 ranks the 10 pairs of questions by the odds ratio computed from the mixed-effect logistic regression model as $exp(\beta + \beta_j)$ in Equation (2). The odds ratio shows to what extent the multiple-choice format of the question is harder than the open-ended format. The bigger the number, the harder the multiple-choice version of the question is. The column Instructor Harder shows a metric we used to measure to what extent instructors think multiple-choice format is harder than the open-ended format. For each pair, if the instructor selects MC to be harder, they get a score of 1; if they select MC and OE are of the same difficulty, they get a score of 0.5; otherwise they get

ID	OE-Score	MC-Score	Odds Ratio (OE/MC)	Student Data Harder	Instructor Harder 1=MC 0=OE
1	0.94	0.7	2.19	MC	0.21
2	0.97	0.81	1.99	MC	0.67
3	0.97	0.84	1.78	MC	0.46
4	0.78	0.71	1.37	Same	0.42
5	0.71	0.7	1.14	Same	0.21
6	0.82	0.83	1.07	Same	0.25
7	0.69	0.7	1.07	Same	0.17
8	0.97	1	1.06	Same	0.33
9	0.92	0.95	1.04	Same	0.17
10	0.92	0.97	0.94	Same	0.21

Table 2: Ranks the 10 pairs of questions by the odds ratio computed from the logistic regression model. Higher odds ratio suggests harder multiple-choice format of the question. Instructor score suggests to which extent instructors predicted the multiple-choice format of the question was harder than the open-ended question.

Instructor \ Student	MC	Same	OE	Total
MC	14	19	0	33
Same	9	19	0	28
OE	31	88	0	119
Total	54	126	0	180

Table 3: This table shows the low alignment between student performance data and instructor prediction data. The number indicates the frequency of instructors (or students) predicting one format of the question to be harder or that the two format are of the same difficulty. The greyed area indicates when instructor prediction and student performance align.

a score of 0. It shows instructors believe the MC format is harder if the score is closer to 1 and vice versa. If student performance data and instructor judgment align, the odds ratio and instructor score columns in Table 2 should rank in the same way. However, this is not what we observe. Additionally, we observed a close to zero correlation between the two columns (Pearson's correlation coefficient = −0.05), suggesting instructor judgment do not align with student performance data. Table 3 displays instructor judgment and student performance data by responses. Each cell indicates how many times the instructor or the student data suggests MC (or OE) is harder. The two greyed cells are where they align. Again, we see that the alignment between student performance and instructor prediction is low.

Following the predictions of relative difficulty of the question pairs, we also asked participants to elaborate on their decision making process. We used a thematic analysis approach to analyze these responses and we summarize three themes below. 1) they followed a trend of predicting open-ended to be harder for reasons such as generating an answer is more challenging and the search space is bigger. 2) they considered multiple-choice to be harder for reasons such as the options are trickier. 3) they asked about how the grading of open-ended is done (how lenient) and needed to take that into consideration. Here are some quotes of responses under each of the three themes.

Instructors said they followed a trend of predicting open-ended questions to be harder for reasons such as "it's hard to generate ideas", "the search space for a good answer is larger", "it's harder when students are asked to recall a terminology compared to multiple-choice questions"

> *"I thought the open-ended questions were harder than the multiple choice questions. This is because it's hard to generate ideas for ways to improve (say) interview questions or interface designs from scratch. The multiple-choice questions model the sorts of things you could think about, and help them get the correct answer more often." – P8*

> *"I selected the multiple choice option as being easier in cases where the student is being asked to recall terminology that I have seen students struggle with remembering. When the question required explanation of a concept or the multiple choice option didn't give significant cues I tended to rate them as similar difficulty or the open-ended easier. Yes, mostly I think open ended questions are harder than multiple choice questions because the search space for a good answer is larger. The multiple choice question restricts what one can think about." – P5*

Instructors also disclosed reasons why they thought multiple-choice questions to be harder or equally hard as open-ended questions, including "when the distractors are difficult", "when multiple-choice do not provide scaffolds", "when multiple-choice also require explanation of concepts", "when multiple-choice ask them to pick the best from several plausible answers".

> *" The criteria I used to decide which problem was harder was to put myself in the shoes of a novice student of the subject. Many times the multiple choice questions restricted potential answers that could be equally plausible, particularly with the earlier questions. I was more biased toward multiple choice questions being harder for students, except when it came to the heuristic evaluation questions. This may have been because the multiple choice questions appeared to provide better scaffolds for the heuristic evaluation than the other scenarios, which were more open ended." – P19*

> *"There is the occasional exception to my trend of always-selecting-open-ended-as-harder. In some cases, when the multiple choice options do not contain a very clear correct answer, multiple choice can be trickier. Many of these questions are subjective, and there might be multiple ways to improve an example. Picking from a group of options that are similarly important can make the multiple choice question trickier in some cases." –P15*

> *"Also I rated any question where the prompt hinged on "most salient" as equally hard because that aspect of the question (rather than the format) would drive my perception of its difficulty." –P3*

In addition, a few participants also raised concerns around how grading is done. For example, "Without providing a rubric for the open-ended items it was hard to tell how they would be scored and thus judge how likely a student would be to get them correct. ", "it depends on how lenient I thought the grading would be on the open-ended questions (i.e., was there one right answer that was hard to recall, or many possibly solutions that students could suggest)". We will discuss instructors' decision making criteria and reasoning in the next section.

7 DISCUSSION

7.1 A Closer Look at Student Responses

The classroom experiments suggest that students get similar performance in matched pairs of multiple-choice and open-ended questions. We took a closer look at student responses to these questions. In Figure 4, we show several example student answers in response to Q1 (Figure 3), including correct and incorrect answers for both question formats. Some exam papers suggest that students are evaluating and comparing options when they work on questions (S2 in Figure 4). Often times, the wrong answer students give in

open-ended questions assemble the incorrect options we present in the MCQs (S5 in Figure 4). And the correct answer students give in open-ended questions also assemble the correct answer in the multiple-choice question (S3 in Figure 4).

Here is a possible explanation for why multiple-choice and open-ended formats are of similar difficulty. The distractors in the MC are based on the common mistakes students have made in the past, i.e., students mistakes resemble distractors in MC, so that the options in MC offer a plausible and not easy search space for students (seen from P5 reasoning). Furthermore, among 4 of the 18 pairs of questions, the multiple-choice version of the question is slightly harder than the matched open-ended version. A possible explanation for this using Q1 (in Figure 3) as an example is that the wrong option B introduces a new error which students would not consider when answering the open-ended question, making it a difficult competitive distractor (seen from P3, P15, P19 reasoning).

7.2 Thinking Elements Required when Answering MC and OE Questions

The results we have seen in the classroom experiments reject Hypothesis 1 and support Hypothesis 2, suggesting that the critical thinking elements required when answering matched well-designed multiple-choice and open-ended questions are learning to evaluate proposed solutions in terms of the deep features that differentiate their correctness. In this section, we delve deeper into the thinking elements required when answering MC and OE questions.

Using Q1 (Figure 3) as an example, when answering the open-ended question, the possible thinking elements required include *(i)* generating candidate solutions and *(ii)* evaluating whether the candidate solution is good or not; when answering the multiple-choice question, the possible thinking elements required is solely *(ii)* evaluating whether the candidate solution is good or not. Our results

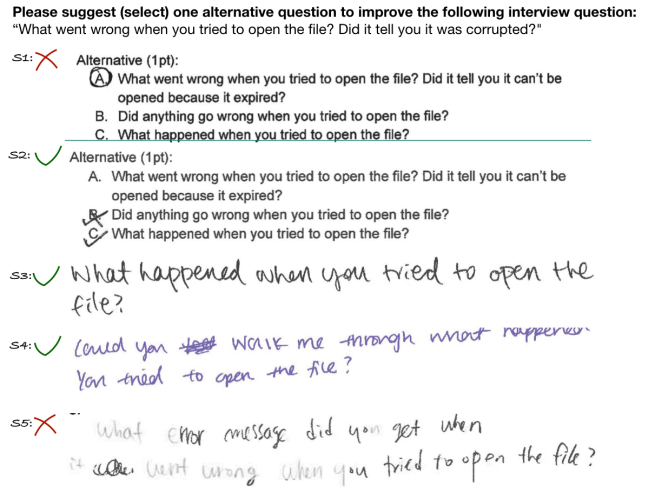

Figure 4: Example student answers in response to Question 1 in Figure 3, including correct and incorrect answers for both question formats (MC and OE).

on student performance data suggest that the thinking elements required in *(i)* generating candidate solutions is insignificant because students displayed similar performance on matched multiple-choice and open-ended questions. Using Q1 as an example, this suggests that when learning to write interview questions, the challenging part is not to come up with an interview question in the first place, but the real challenge lies in evaluating whether a candidate interview question is good or not, i.e., whether it is leading, asking a yes/no question, asking secondhand information, etc. In addition to designing interview questions, we have found several other HCI knowledge components where the generation efforts required are insignificant.

The practical implication for this research is to encourage instructors to examine and gauge the relative generation and evaluation efforts required in problem-solving before giving students open-ended tasks with an emphasis on generation new content. For topics that require a significant amount of evaluation efforts, evaluation-type exercises such as multiple-choice questions can be efficient for student learning since they exercises the critical thinking elements such as evaluating the quality of candidate solutions. From a practical standpoint, evaluation-type exercises such as multiple-choice questions are much easier to scale, offer real time feedback and can enable repeated practice in varying contexts for students.

7.3 Implication for HCI Education and Practice

We consider our work as complementary to the existing research on HCI education and pedagogy. To make it clear, we do not think students can become good UX designers or researchers with multiple-choice practice only. We value studio-based and project-based learning approaches and instructional techniques that offer students opportunities to engage with real users and manage authentic projects [18, 26, 34, 37, 51]. As HCI educators ourselves, we surface an issue in HCI education that emphasizes content generation and project-based learning even when students are not ready. As examples, we hear frequent comments from instructors such as "Students are asked to design a survey when they didn't actually know how to design a survey. Many assignments turned in were in very bad shape and I had to tell the students to go back and redo it." Similar challenges have been reported in recent work such as [34]. At the same time, we have also surfaced a negative sentiment towards the use of multiple-choice questions, as shown in our survey on HCI education and also in the literature on other topics [30]. Although it is generally believed that multiple-choice questions target lower Bloom goals [5] such as "Remembering" and "Understanding", we see in our cases, multiple-choice questions target higher Bloom goals, such as "Evaluation." The multiple-choice questions used in our experiments encourage students to compare and evaluate the candidate solutions and decide on better solutions to specific problem scenarios, which does not assume one definite fact answer to the questions. Our work suggests that giving students opportunities to learn and practice HCI skills through evaluation-type activities would be valuable before engaging students with complex project-based learning. The practical benefits are that the evaluation-type activities, such as multiple-choice questions are much easier to

scale, offer real time feedback and can enable repeated practice in varying contexts for students.

We also need to note and clarify here that the HCI topics we tested in our experiments are all frequently used techniques (interviews, think-aloud, heuristic evaluation, etc) which are better defined than some other HCI techniques. For example, there is a general agreement among HCI researchers on what are considered as good interview questions and what are good think-aloud protocols. For techniques that are less well-defined, we consider such evaluation-type exercises, i.e., which would be a better candidate solution can help define and improve the curricula on those topics.

7.4 Expert Blindspot and Instructional Design

In the instructor survey, many instructors revealed that they made the judgments based on the assumption that recognition is easier than recall, which makes the multiple-choice questions easier than their open-ended counterparts. After we showed instructors multiple-choice questions that could target higher Bloom goals, one instructor commented that they didn't like to use multiple-choice questions because "the ones I write are bad, and i haven't mastered how to write good ones." This might be a reason that reinforces the negative sentiment towards using multiple-choice questions for learning.

Other instructors had made judgements based on the reasoning that open-ended questions exercise critical thinking and multiple-choice ones do not. Our analysis shows that the thinking elements required in answering these questions may not align with instructors' hypothesis and prediction. Some instructors mentioned that they made the judgments based on how hard they thought the distractors were. When distractors seemed trickier or there were multiple options that could be correct, they thought the multiple-choice question could be harder. Although it was true that competitive distractors could make a multiple-choice question harder, it appeared that instructors were not very effective in identifying which questions had competitive distractors. For example, Q1 in Figure 3 has the highest odds ratio among all questions and the distractors are very competitive. However, 75% of the instructors thought the open-ended version would be harder. We consider the above as reasons for the expert blind spots we have observed when experts are predicting the knowledge gaps of novices when learning HCI.

As HCI educators ourselves, we share these blind spots with many of our participants. The message here is constructive rather than critical. To combat expert blind spots, our work suggests that reasoning behind educational decisions can be probed through well-designed, low-effort, experimental comparisons toward more nuanced and accurate reasoning and decision making, and ultimately better design.

This work also offers suggestions to establish the profession of Learning Experience (LX) designers to develop curriculum in higher education. In many other domains, design of products to support the workflow of professionals require expertise from both domain experts and interaction designers, e.g., interaction designers design products to support doctors' decision making [57]. However, instructors are frequently required to take on both roles though their expertise does not prepare them for both. Our work suggests that, consistent with other design practices, to improve quality of learning design in higher-education, establishing roles such as learning designers or learning engineers is desirable. We demonstrate well-designed, low-effort, experimental comparison techniques that would allow LX designers to discover and employ empirically-rooted instructional and assessment methods. When designing learning experience, LX designers need to focus more on the underlying cognitive processes being measured instead of the format or surface features of the tasks [46]. When faced with the choice of using either MCQs or open-ended questions, it is important for LX designers to consider the nature of the learning objectives, i.e., the relative difficulty of the generation and evaluation processes involved. For content domains where evaluating candidate solutions could be challenging and worthwhile, such as the domains we have tested, there is more promise and benefit of using MCQs for scalable and high quality instruction and assessment.

7.5 Online Learning for Scale and Quality

In this work, we investigate the relative benefits of multiple-choice and open-ended questions, situated in HCI research methods. We indicate more promise than concern in using MCQs for scalable instruction and assessment in at least some HCI domains. With the recent development of learning technologies, we envision the appropriate use of high quality multiple-choice questions that target higher Bloom goals could help us achieve scale without sacrificing learning quality in some domains. Prior work has shown that high quality multiple-choice questions can be semi-automatically produced using a learnersourcing approach [52]. With recent advances in AI-based automatic question generation techniques, such content creation process can be further scaled and expedited. In addition, multiple-choice questions can be presented to students in an adaptive order leveraging adaptive problem selection techniques such as intelligent tutoring systems [22, 39]. Furthermore, multiple-choice questions may be integrated into conversational agents and pedagogical agents that give students practice opportunities as they explore an open-ended task [49, 53, 56].

This work offers theoretical understanding and empirical evidence on when, why, and how multiple-choice questions can be of high quality, i.e., exercising critical thinking elements instead of exercising purely recognition. This contributes to the HCI and learning technology design community in employing multiple-choice questions in learning systems, and developing techniques to produce high quality evaluation-type exercises that could achieve quality learning at scale.

8 CONCLUSION

First, this paper indicates more promise than concern in using MCQs for scalable instruction and assessment, with the goal of providing high quality education to more and more learners through online or physical programs. We demonstrate a experimental comparison technique that can be employed to compare alternative instructional and assessment methods, with the goal of designing learning experience that are both scalable and high quality. Second, this paper provides further evidence that expert blind spots exist, we observe that instructor intuition and reasoning sometimes do not match those of student performance. When considering

learning experience design, a deeper analysis of the underlying cognitive processes students would engage in is desired. Finally, faculties often need to act as both domain experts and LX designers in many higher-education contexts, with limited time, resources and preparation for the dual roles. We recommend to establish the profession of Learning Experience (LX) designers, whose work can support the instructional design and development in higher education, and also contribute to the broader HCI interaction design practices.

ACKNOWLEDGMENTS

This work was funded by NSF grant ACI-1443068 and DRL-1740798. In addition, we thank Raelin Musuraca, Amy Ogan, Jason Hong, Steven Dang, Mark Guzdial, Scott Klemmer, Anhong Guo and all participants.

REFERENCES

[1] José Abdelnour-Nocera, Mario Michaelides, Ann Austin, and Sunila Modi. 2012. An intercultural study of HCI education experience and representation. In *Proceedings of the 4th international conference on Intercultural Collaboration*. 157–160.
[2] Piotr D Adamczyk and Michael B Twidale. 2007. Supporting multidisciplinary collaboration: requirements from novel HCI education. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1073–1076.
[3] Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2016. Ontology-based multiple choice question generation. *KI-Künstliche Intelligenz* 30, 2 (2016), 183–188.
[4] Douglas Bates, Deepayan Sarkar, Maintainer Douglas Bates, and L Matrix. 2007. The lme4 package. *R package version* 2, 1 (2007), 74.
[5] TAXONOMY MADE EASY BLOOM'S. 1965. *Bloom's taxonomy of educational objectives*. Longman.
[6] Clodis Boscarioli, Luciana AM Zaina, Sílvia Amélia Bim, Simone Diniz Junqueira Barbosa, and Milene S Silveira. 2016. HCI Education in Brazil from the Results of the Workshop on Teaching of HCI. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*. 1–4.
[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[8] Neil CC Brown and Amjad Altadmri. 2014. Investigating novice programming mistakes: Educator beliefs vs. student data. In *Proceedings of the tenth annual conference on International computing education research*. 43–50.
[9] Neil CC Brown and Amjad Altadmri. 2017. Novice Java programming mistakes: Large-scale data vs. educator beliefs. *ACM Transactions on Computing Education (TOCE)* 17, 2 (2017), 1–21.
[10] Elizabeth Churchill, Jennifer Preece, and Anne Bowser. 2014. Developing a living HCI curriculum to support a global community. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 135–138.
[11] Elizabeth F Churchill, Anne Bowser, and Jennifer Preece. 2016. The future of HCI education: a flexible, global, living curriculum. *interactions* 23, 2 (2016), 70–73.
[12] Jason Day and Jim Foley. 2006. Evaluating web lectures: A case study from HCI. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. 195–200.
[13] Paul Denny, Andrew Luxton-Reilly, and John Hamer. 2008. The PeerWise system of student contributed assessment questions. In *Proceedings of the tenth conference on Australasian computing education-Volume 78*. Citeseer, 69–74.
[14] Louis Deslauriers, Logan S McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. 2019. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences* (2019), 201821936.
[15] Barbara J Ericson, Lauren E Margulieux, and Jochen Rick. 2017. Solving parsons problems versus fixing and writing code. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research*. 20–29.
[16] Chase Felker, Radka Slamova, and Janet Davis. 2012. Integrating UX with scrum in an undergraduate software development project. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*. 301–306.
[17] Steven C Funk and K Laurie Dickson. 2011. Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology* 38, 4 (2011), 273–277.
[18] Colin M Gray. 2014. Evolution of design competence in UX practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1645–1654.
[19] Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 2016. Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 458–469.
[20] Instructure, Inc. 2019. Canvas. https://www.canvaslms.com/research-education
[21] David A Joyner, Wade Ashby, Liam Irish, Yeeling Lam, Jacob Langston, Isabel Lupiani, Mike Lustig, Paige Pettoruto, Dana Sheahen, Angela Smiley, et al. 2016. Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 399–408.
[22] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. 1997. Intelligent tutoring goes to school in the big city. (1997).
[23] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.
[24] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale*. ACM, 111–120.
[25] Kenneth R Koedinger and Mitchell J Nathan. 2004. The real story behind story problems: Effects of representations on quantitative reasoning. *The journal of the learning sciences* 13, 2 (2004), 129–164.
[26] Panayiotis Koutsabasis, Spyros Vosinakis, Modestos Stavrakis, and Panagiotis Kyriakoulakos. 2018. Teaching HCI with a studio approach: Lessons learnt. In *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*. 282–287.
[27] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 75–84.
[28] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education* 30, 1 (2020), 121–204.
[29] Jeri L Little and Elizabeth Ligon Bjork. 2015. Optimizing multiple-choice tests as tools for learning. *Memory & Cognition* 43, 1 (2015), 14–26.
[30] Jeri L Little, Elizabeth Ligon Bjork, Robert A Bjork, and Genna Angello. 2012. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological science* 23, 11 (2012), 1337–1344.
[31] Mitchell J Nathan and Kenneth R Koedinger. 2000. An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction* 18, 2 (2000), 209–237.
[32] Mitchell J Nathan and Kenneth R Koedinger. 2000. Teachers' and researchers' beliefs about the development of algebraic reasoning. *Journal for Research in Mathematics Education* (2000), 168–190.
[33] Mitchell J Nathan and Anthony Petrosino. 2003. Expert blind spot among preservice teachers. *American educational research journal* 40, 4 (2003), 905–928.
[34] Alannah Oleson, Meron Solomon, and Amy J Ko. 2020. Computing Students' Learning Difficulties in HCI Education. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[35] Leo Porter, Cynthia Taylor, and Kevin C Webb. 2014. Leveraging open source principles for flexible concept inventory development. In *Proceedings of the 2014 conference on Innovation & technology in computer science education*. 243–248.
[36] Henry L Roediger III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255.
[37] Wendy Roldan, Xin Gao, Allison Marie Hishikawa, Tiffany Ku, Ziyue Li, Echo Zhang, Jon E Froehlich, and Jason Yip. 2020. Opportunities and Challenges in Involving Users in Project-Based HCI Education. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[38] James M Royer, Cheryl A Cisero, and Maria S Carlo. 1993. Techniques and procedures for assessing cognitive skills. *Review of Educational Research* 63, 2 (1993), 201–243.
[39] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[40] Mehdi SM Sajjadi, Morteza Alamgir, and Ulrike von Luxburg. 2016. Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 369–378.
[41] Eunice Sari and Bimlesh Wadhwa. 2015. Understanding HCI education across Asia-Pacific. In *Proceedings of the International HCI and UX Conference in Indonesia*. 65–68.
[42] Richard A Schmidt and Robert A Bjork. 1992. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science* 3, 4 (1992), 207–218.
[43] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. 2017. Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. ACM, 81–88.
[44] Megan A Smith and Jeffrey D Karpicke. 2014. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory* 22, 7 (2014), 784–802.

[45] Olivier St-Cyr, Craig M MacDonald, Elizabeth F Churchill, Jenny J Preece, and Anna Bowser. 2018. Developing a community of practice to support global HCI education. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.

[46] Brenda Sugrue. 1995. A Theory-Based Framework for Assessing Domainl-Specific Problem-Solving Ability. *Educational Measurement: Issues and Practice* 14, 3 (1995), 29–35.

[47] Brenda Sugrue, Noreen Webb, and Jonah Schlackman. 1998. The Interchangeability of Assessment Methods in Science. CSE Technical Report 474. (1998).

[48] Jennyfer Lawrence Taylor, Jessica Tsimeris, XuanYing Zhu, Duncan Stevenson, and Tom Gedeon. 2015. Observations from teaching HCI to Chinese students in Australia. In *Proceedings of the ASEAN CHI Symposium'15*. 31–35.

[49] Gaurav Singh Tomar, Sreecharan Sankaranarayanan, Xu Wang, and Carolyn Penstein Rosé. 2017. Coordinating collaborative chat in massive open online courses. *arXiv preprint arXiv:1704.05543* (2017).

[50] Anna Vasilchenko, Adriana Wilde, Stephen Snow, Madeline Balaam, and Marie Devlin. 2018. Video coursework: opportunity and challenge for HCI education. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*. 1–3.

[51] Mihaela Vorvoreanu, Colin M Gray, Paul Parsons, and Nancy Rasche. 2017. Advancing UX education: A model for integrated studio pedagogy. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1441–1446.

[52] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing Student Open-Ended Solutions to Create Scalable Learning Opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (Chicago, IL, USA) *(L@S '19)*. ACM, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/3330430.3333614

[53] Xu Wang, Miaomiao Wen, and Carolyn Rose. 2017. Contrasting explicit and implicit support for transactive exchange in team oriented project based learning. Philadelphia, PA: International Society of the Learning Sciences.

[54] Noreen M Webb, Jonah Schlackman, and Brenda Sugrue. 2000. The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education* 13, 3 (2000), 277–301.

[55] Lauren Wilcox, Betsy DiSalvo, Dick Henneman, and Qiaosi Wang. 2019. Design in the hci classroom: Setting a research agenda. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 871–883.

[56] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[57] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4477–4488.

[58] Nesra Yannier, Kenneth R Koedinger, and Scott E Hudson. 2015. Learning from mixed-reality games: Is shaking a tablet as effective as physical observation?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1045–1054.