

Harnessing Student Solutions to Support Learning at Scale

CMU-HCII-20-106
August 2020

Xu Wang

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA 15213
xuwang@cs.cmu.edu

Thesis Committee:

Kenneth R. Koedinger (Co-chair), HCII, CMU
Carolyn Penstein Rose (Co-chair), LTI and HCII, CMU
Jeffrey P. Bigham, HCII and LTI, CMU
Chinmay Kulkarni, HCII, CMU
Scott Klemmer, University of California San Diego

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Copyright © 2020 Xu Wang

Keywords: Deliberate practice, immediate feedback, human-machine collaboration, artificial intelligence in education, learning at scale, higher education, open-ended assignment, multiple-choice question, automatic question generation, learnersourcing, crowdsourcing

To Mom, Dad, and Anhong.

Abstract

A challenge to meet the demand on higher education and professional development is to scale these educational opportunities while maintaining their quality. My dissertation work tackles this challenge by harnessing examples from existing resources to enable the creation of scalable and quality educational experiences. Deliberate practice targeting specific skills, appropriate scaffolding with timely feedback helps novices become experts. However, the feature of deliberate practice with timely feedback is often missing in college instruction. On the one hand, instructors believe they should assign flexible work, but the very multifacetedness that makes it authentic impedes students' ability to learn because they rarely get timely, attribute-specific feedback. On the other hand, instructors find designing materials that offer focused practice and immediate feedback to be time-consuming and challenging.

This dissertation contributes insights about developing effective learning at scale systems by leveraging the complementary strengths from peers, experts, and machine intelligence, differentiating it from existing systems that solely rely on machine or crowds of peers. This dissertation introduces a technique UpGrade, which uses student solution examples to semi-automatically generate multiple-choice questions for deliberate practice of higher order thinking in varying contexts. From experiments in authentic college classrooms, I show that UpGrade helps students gain conceptual understanding more efficiently and helps improve students' authentic task performance. Through an iterative design process with instructors, I demonstrate the generalizability of this approach and offer suggestions to improve the quality and efficiency of college instruction.

This dissertation suggests another layer to further distinguish knowledge components, by the required generation and evaluation efforts in problem-solving. The practical implication for a more nuanced understanding of knowledge components is to help instructors make more nuanced and accurate instructional decisions, e.g., using "evaluation-type" exercises for evaluation-heavy skills. This dissertation provides further evidence that instructors have so-called "expert blind spots", revealed through cases where their beliefs and student performance do not match. More generally, this work suggests that the 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.

Acknowledgments

I'd like to acknowledge all the amazing guidance, support, friendship and care I've received over the past 6 years. I wouldn't be here today without any of you.

I'd like to thank my advisors Carolyn Rose and Ken Koedinger for showing me how to do research and for the incredible guidance and support over the years. The things I have learnt from them are far beyond research-related knowledge and skills. I started working with Carolyn my first year at CMU, and her dedication and enthusiasm about research has influenced me since day 1. I have also become a lot better at time management and developed better self-discipline during my Ph.D, which I got huge influence from Carolyn. I started working with Ken in the Spring semester of my third year in the program. Ken has shown me how to be a great mentor. He always listens and is willing to understand my points even when they are awkwardly framed. Ken has always encouraged me to think behind the scenes, so that I don't stop at making something to work, but continue to ask questions such as why it works like that. I couldn't be luckier to have been able to work with both of them, and I can only strive to relay the incredible mentorship I have received to my students in the future.

I want to thank my committee members Scott Klemmer, Chinmay Kulkarni and Jeff Bigham. Their questions greatly helped me consolidate and articulate the contributions of this dissertation. Even though we only had limited interactions, there are so many great insights coming out of them.

I'd like to thank Tovi Grossman and Justin Reich, who have been great mentors to me, helped me see diverse ways of doing research, and helped me explore new research directions in software learning and teacher education.

I'd like to thank all my wonderful collaborators, Raelin Musuraca, Sree Sankaranarayanan, Miaomiao Wen, Meredith Thompson, Ben Lafreniere, Gaurav Tomar, Chris Bogart, Dan Roy, Majd Sakr, Diyi Yang, Keith Maki, Amanda Godley, and the incredible undergraduate and masters students I had the honor to mentor and collaborate with, including Teja Talluri, Yali Chen, Kexin Yang, Alexis Soto, Jiaojiao Song, and Rajitha Pulivarthy. This dissertation wouldn't have been possible without all your input and help.

I have also received so much support from mentors, colleagues and friends in the past 6 years. I'd like to thank faculties in HCII and PIER who have offered help to my research, Sharon Carver, Ellen Ayoob, Amy Ogan, Bob Kraut, Jim Morris, David Klahr, Jason Hong, and Skip Shelly.

I'd like to thank the amazing colleagues and friends who I had the luck to know and interact with. Thanks for helping pilot my study, answering my questions, giving feedback to my talks, and being there for me. I can't be grateful enough to all the support and friendship I have received, and I have been so inspired by every one of you, Shikun Zhang and Zheng Yao (for the chitchat and for being there for me all the time), Anna Kasunic (for starting and exploring the Ph.D. together, for helping with my English, and for being there for me), Qian Yang (for always answering my random questions about design), Francesca Xhaja, Steven Dang, Paulo Carvalho, Nesra Yannier, Nick Diana, Julia Cambre, Haojian Jin, Yohan Jo, Michael Yoder, Qinlan Shen, Zhengzhong Liu, Xu Zeng, Julia Qian, Toby Li, Xieyang Liu, Tianying Chen, Yang Zhang, Judith Uchidiuno, Siyan Zhao, Nathan Hahn, Judy Choi, Kristin Williams, Michael Madaio, Ken Holstein, Yasmine Kotturi, Moe Alburaiqi, Nick Lewis, Elizabeth Onstwedder, Ting Han, Yijing Cui, Felicity Fu, Yuxin Qi, Iris Howley, Nikola Banovic, Haiyi Zhu, Hong Shen, Yanjin Long, Ruogu Kang, Keyang Xu, Zi Yang, Ziyun Deng, Barbara Ericsson, Su Cai, Jo Bodnar, Michael Bett, Audrey Russo, Queenie Kravitz, and many others!

I'd like to thank my best friend and husband Anhong Guo, who I met through the Ph.D. program. Thanks for giving me all the courage I need to get here today, and I feel extremely lucky to have been able to learn from you everyday.

Finally, I'd like to thank my parents Fengxia Sun and Dehai Wang, for their unconditional love and support, for always believing in me and for everything they have done for me.

Contents

1	Introduction	1
2	Related Work	9
2.1	Automatic Question Generation for Educational Purposes	9
2.2	Learnersourcing	10
2.3	Expertise Development	11
2.3.1	Deliberate Practice	11
2.3.2	Constraint-based Expertise Acquisition	11
2.4	The Knowledge-Learning-Instruction Framework	12
2.5	Problem Solving, Worked Examples, Feedback	13
2.6	Quality Control through Psychometric Approaches	13
3	UpGrade: A Learnersourcing Technique	15
3.1	Introduction	15
3.2	Related Work	16
3.2.1	Learnersourcing Techniques	17
3.2.2	Worked Examples and Scaffolding	17
3.2.3	Repeated Practice and Feedback	18
3.2.4	Quality Control Methods	18
3.3	Formative Study: Assignment Survey	18
3.4	UpGrade	19
3.4.1	Solution Logging	21
3.4.2	Solution Segmentation Based on Assignment Rubric	21
3.4.3	Question Creation	23
3.5	Classroom Experiment of UpGrade	27
3.5.1	Crossover Experiment Design	27
3.5.2	Learning Outcome Measure	28
3.6	Experiment Results	29
3.6.1	User Experience and Feedback	30
3.7	Discussion, Limitations and Future Work	31
3.7.1	Structured Text Data Logging	31
3.7.2	UpGrade As A Primer To Open-ended Assignment	31
3.7.3	Quality Control and Quality Enhancement	31
3.8	Conclusion	32

4	Evaluation of UpGrade Using Open-ended Task Performance Measures	33
4.1	Study Design	33
4.2	Study Materials	34
4.2.1	Implementation of UpGrade	36
4.2.2	Analysis Methods	36
4.3	Experiment Results	44
5	Iterative Design of QuizMaker	47
5.1	Initial formative interviews	47
5.1.1	Methods	47
5.1.2	Themes emerging from the formative interview study	47
5.2	Iterative design, Co-design of QuizMaker	49
6	Using Psychometric Methods to Support Automatic Item Generation and Evaluation	53
6.1	Cronbach’s Alpha to Evaluate Consistency	53
6.2	MTurk Study	54
6.2.1	Participants and Procedure	54
6.2.2	Prune Out Unreliable Question Items	54
6.2.3	Question Face Validity Inspection	54
6.2.4	Cost-effectiveness of Quality Control	55
6.3	Summary	56
7	Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality	59
7.1	Introduction	59
7.2	Assessment Comparisons to Indicate Learning Benefits	61
7.3	Related Work	62
7.3.1	Debate Around the Use of Open-ended vs MC Questions	62
7.3.2	Importance of Instructor Belief	63
7.3.3	Relative Difficulty of Matched Questions	63
7.4	Methods	64
7.4.1	Hypotheses about Students’ Underlying Cognitive Processes When Answering Questions	64
7.4.2	Design of Matched Pairs of Questions	65
7.5	Instructor Belief Survey	66
7.5.1	Participants	66
7.5.2	General Beliefs	66
7.5.3	Predictions on Relative Difficulty of Matched Questions	67
7.5.4	Instructor Reasoning	67
7.6	Classroom Experiments	68
7.6.1	Study Design and Implementation	68
7.6.2	Answer Grading and Dataset	68
7.6.3	Multiple-choice Questions Do Not Avoid the Hard Part	69
7.6.4	Instructor Reasoning and Student Data Conflicts	70

7.6.5	Answer Examples	72
7.7	Discussion	72
7.8	Conclusion	73
8	Practice-Based Teacher Questioning Strategy Training with ELK: A Role-Playing Simulation for Eliciting Learner Knowledge	75
8.1	Introduction	75
8.2	Related Work	77
8.2.1	Teacher Classroom Discourse Training and Support	78
8.2.2	Role-play and Simulation Systems for Learning	78
8.2.3	How to Elicit Learner Knowledge	78
8.2.4	Learnersourcing and the Benefits of Solution Evaluation	79
8.3	ELK: A Role-Playing Simulation System	79
8.3.1	Text-based Role-Play	81
8.3.2	Coding Activity	82
8.4	Evaluation of ELK	82
8.4.1	Participants	82
8.4.2	Study Components	83
8.4.3	Procedure	84
8.4.4	Outcome Measures	84
8.4.5	Survey Analysis Method	86
8.5	Results	87
8.5.1	Effectiveness of ELK	87
8.5.2	“Coding” is an Effective Supplementary Activity in ELK	90
8.5.3	User Experiences, Challenges and Feedback	92
8.6	Design Implications	93
8.6.1	Content Domain Knowledge is Essential	93
8.6.2	Role-Play Helps Participants Gain Perspectives of Relevant Stakeholders	93
8.6.3	Focused Practice through Evaluation Activities	94
8.6.4	Different Modalities of Role-Play	94
8.7	Conclusion	94
8.8	Appendix	94
9	Leveraging Community-Generated Videos and Command Logs to Classify and Recommend Software Workflows	97
9.1	Introduction	97
9.2	Related Work	98
9.2.1	Relevant Applications for Software Learning	98
9.2.2	User Data Mining	99
9.3	Dataset	100
9.4	Understanding User Tasks from Videos	100
9.4.1	Frequent Pattern Mining Approach	100
9.4.2	Topic Modeling Approach	101
9.5	Study 1: Task Categorization Validation	104

9.5.1	Video Study Design	104
9.5.2	Quantitative Results and Analysis	107
9.5.3	Qualitative Analysis of User Feedback	108
9.6	Hierarchical Task Identification	109
9.7	Recommender System	111
9.7.1	Recommender System Design Space	111
9.7.2	Recommendation Algorithms	112
9.8	Study 2: Workflow Recommendations	114
9.8.1	Participants and Procedure	114
9.8.2	Quantitative Results and Analysis	114
9.8.3	Qualitative Analysis of User Feedback	115
9.9	Discussion and Future Work	115
9.9.1	Incorporating Heuristics Based on Expertise Levels	116
9.9.2	User In-the-Loop Recommender Systems	116
9.9.3	Generalizability	117
9.9.4	Limitations	117
9.10	Conclusion	117
10	List of Contributions	119
10.1	Human-Computer Interaction and Learning Technologies	119
10.2	Learning Sciences	121
11	Discussions and Future Work	125
11.1	Creating and Providing Deliberate Learning Opportunities	125
11.1.1	Generalizability - Active learnersourcing	125
11.1.2	Scalability - Data sharing and reuse	126
11.1.3	Varying inputs, outputs and domains	128
11.1.4	Social Learning	128
11.1.5	Adaptivity	128
11.1.6	Machine Learning Enhanced Quality Control	129
11.1.7	Potential Risks and Ways to Mitigate Them	129
11.2	Artificial Intelligence in Education	129
11.3	Human-AI Collaboration	131
11.4	Instruction and Assessment in Higher Education	132
11.4.1	Scaffold before Open-ended Work	132
11.4.2	Discrepancy between Practice and Theories	132
11.4.3	Service Design in Higher Education	132
11.5	Professional Training and Informal Learning	133
11.6	Generation vs. Evaluation, Creativity	133
	Bibliography	135

List of Figures

- 1.1 Five Step Workflow of UpGrade. Step 1 almost happens naturally. Past students’ written solutions are logged, they can be logged in PDF formats or through online forms. In step 2, the system segments the written solutions into components based on the assignment template (for example, through keyword extraction). This step requires a manual check to make sure the data is meaningfully segmented. In Step 3, we ask the instructor to specify a question creation schema, in which the instructor will specify which components in the source will be used in the target question. In step 4, the system reorganizes the source data based on the schema and leverages Natural Language Processing (NLP) techniques to create multiple-choice questions. In the last step, we ask instructors to select questions from the large question pool. 3
- 1.2 QuizMaker Interface: Segmented past student work is displayed to the left; Instructors can create question schemata and preview the question on the right. . . 4
- 1.3 UpGrade takes advantages of the strengths of peers, machine and instructors. . . 5
- 1.4 In this case, peers who can be users of the software everywhere in the world contribute demonstration videos. We collect these videos (and their command logs) as input and use machine to identify an end user’s workflow, expert input is elicited in this process. With this approach, existing online repositories may be repurposed as targeted tutorials for end users 7

- 3.1 Problem-solving assignment classification from 6 HCI courses. 20
- 3.2 UpGrade’s workflow. 21
- 3.3 Rubric items for open-ended assignments on the topic of Survey Design and Heuristic Evaluation. 22
- 3.4 Four components used in UpGrade question schemata. 22
- 3.5 Example instantiations of the three UpGrade question schemata: (a) Question-Answer, (b) Question-Answer-Explanation, and (c) Answer-Feedback. 23
- 3.6 An example question created by UpGrade using the Question-Answer schema. 25
- 3.7 An example question created by UpGrade using the Question-Answer-Explanation schema. 25
- 3.8 An example question created by UpGrade using the Question-Answer-Explanation schema (Revision variation). 26
- 3.9 An example question created by UpGrade using the Answer-Feedback schema. 26
- 3.10 Student average quiz score in percentage by condition and content with standard error bars. 29

3.11	Student average assignment completion time in hours by condition and content with standard error bars.	30
4.1	Study 2 Design:	34
4.2	Study 2 Materials: Assignment Announcement for “Ideation, Storyboarding and Speed Dating ”	35
4.3	Study 2 Materials: “Ideation, Storyboarding and Speed Dating” Evaluation Component Prompt	35
4.4	Study 2 Materials: “Ideation, Storyboarding and Speed Dating” Writing Component Prompt	36
4.5	Example question of Section 1: asking students to match a user need for a shown storyboard	37
4.6	Example question of Section 2: asking students to match a lead question for a shown storyboard	38
4.7	Example question of Section 3: asking students to evaluate whether the set of three storyboards follow a progression of riskiness. The feedback here is written by the instructor post-hoc, in step 5 of the workflow (shown in Figure 4.9	39
4.8	Example question of Section 4: asking students to evaluate the quality of a shown storyboard, through matching storyboards with previous instructor feedback.	40
4.9	Five Step Workflow of UpGrade	41
4.10	Example of segmentation: through a keyword matching algorithm, the solution on the left is transformed into segmented responses on the right.	41
4.11	UFT Student Work Example 1	42
4.12	UFT Student Work Example 2	42
4.13	Storyboard Student Work Example 1 (Descriptions of the storyboard and relevant meta-data for speed dating.	43
4.14	Storyboard Student Work Example 1	43
4.15	Storyboard Student Work Example 2	44
4.16	Average Percentage Score for both conditions on the two topics.	45
5.1	Five Step Workflow of UpGrade	50
5.2	Screenshot of a prototype system: users see a mapping between the original submission and the segmented solutions.	51
5.3	Segmented past student work is displayed to the left; Instructors can create question schemata and preview the question on the right	52
6.1	Average accuracy in detecting 19 reliable items and 11 unreliable items on different crowd size (across 100 iterations).	56
6.2	Accuracy in detecting the X least reliable items (X in the range of 1-11) varied by crowd sizes (across 100 iterations).	56
7.1	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.	64

7.2	Survey questions that ask about instructors general belief about using MCQs and open-ended questions in their teaching.	66
7.3	Example student answers in response to question 1 in Figure 7.1, including correct and incorrect answers for both formats.	71
8.1	“Role-play” interface in ELK for the “Student” player. The profile is shown on the left, including the student’s (mis)conceptions about the topic. Players chat on the right.	80
8.2	Teacher and student profiles on the topic of Grade 6 Algebra	80
8.3	Teacher and student profiles on the topic of Heredity	81
8.4	“Coding” activity in ELK. The user reads an authentic‘ transcript generated from past role-play sessions and assign a questioning move to each line in the transcript.	81
8.5	The study spans two 90-minute class meetings. In the first class meeting, participants play three rounds of ELK without feedback. This means for the “Role-play” activity, the quiz is disabled, and for the “Coding” activity, participants do not get feedback after making a selection. Participants are paired and assigned to one of the two conditions. The only difference is whether they complete the “Role-play” or the “Coding” activity in the second round. In the second class meeting, participants play three rounds of ELK with feedback. Participants remain in the same pair and switch roles. The pairs that did the “Coding” activity in Round 2 will now do the “Role-play” activity in Round 5 and vice versa.	83
8.6	Participants increased positive questioning moves from Round 1 to Round 3 across conditions.	88
8.7	Left: excerpts from Round 1 of P68 playing the “Teacher” role. The participant used multiple “Telling” moves and directly told the student the correct answer when the student made a mistake. Right: excerpts from Round 3 of P68 playing “Teacher” role. When the student gave an unexpected answer, P68 used “Eliciting” and “Probing” moves to understand the thinking process of the student.	89
9.1	Task names, as labeled by experts, with the number of videos that are categorized for each task.	103
9.2	Expert ratings for Q1:“How frequently/likely do you think these commands would be used together”	104
9.3	Example labeling question – Select the category that best describes the task being performed in the video.	105
9.4	Similarity question – Rate the similarity of the tasks performed in the two videos.	106
9.5	Similarity between a video and all other videos. We selected similar videos from the green shaded area, and dissimilar videos from the red shaded area.	106
9.6	Summary of results for Study 1, grouped by video set.	108
9.7	Summary of agreement between the algorithm and participants’ responses to the labeling questions.	108
9.8	Distribution of patterns by task.	110
9.9	Example task distribution, user vs. similar users.	112
9.10	Summary of our hierarchical approach.	116

11.1 Student solution example on the topic of usability findings. The severity of the problem is further broken down into three aspects. 126

11.2 The active learnersourcing practice we piloted in UCRE. 127

11.3 An example UpGrade-created question with active learnersourcing. 127

11.4 Arrows #1 and #2 are existing focuses of AI in education; #3 and #4 indicate my contributions. #1 stands for direct AI support to learners, e.g., automatic grading, chat bots. #2 stands for techniques that intervene during instruction, e.g., intelligent tutoring systems that support adaptive question selection. #3 stands for techniques to support the authoring of content leveraging past student data. #4 stands for techniques to support interaction between peers. 130

List of Tables

3.1	Examples of “evaluation-heavy” problem-solving skills.	20
3.2	A data excerpt produced by the segmentation step: past assignment solutions were segmented into sections based on the assignment rubric. Instructor and peer feedback was associated with solution segments when available.	23
3.3	Course instructor specified a question creation schema for each rubric item in the assignment.	27
3.4	Space: number of questions created in total; Pool: number of questions used in the experiment; Trial: number of questions presented to students in each trial. . .	28
4.1	Parameter estimates and p-value for the repeated measures regression analysis at the team level and at the individual level	46
6.1	The Pearson’s correlation of each question item with the total score and the average Cronbach’s alpha for the set when the item is dropped. Higher correlation and lower Cronbach’s alpha indicates higher reliability.	55
7.1	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.	60
7.2	Ranks the 10 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.	70
7.3	This table shows counts of observations where instructor predicted this question format to be harder and observations where student performance data suggests this question format to be harder. The greyed area shows when the two aligns. . .	71
8.1	Brief version of the coding manual, with definitions and examples of the 5 questioning moves.	86
8.2	Distribution of participants across conditions	87
8.3	Parameter estimates and p-value for both repeated-measures linear regression models comparing the effectiveness of “Coding” and “Role-play” activities. Both models show that there is no observed difference between the two conditions. . .	91
8.4	Full transcript of an example role-play session (1)	95

8.5	Full transcript of an example role-play session (2)	96
9.1	A design space for workflow recommender systems.	112
9.2	Study 2 results: users' rating on the relevance and familiarity of the recommended tutorial videos.	115

Chapter 1

Introduction

A challenge to meet the demand on higher education and professional development is to scale these educational opportunities while maintaining their quality. Traditionally, teachers or experts take the most responsibility in providing learning opportunities to students, e.g., giving lectures, offering feedback. However, the efforts required from experts (e.g., offering feedback to students' open-ended work) make such learning opportunities less scalable. AI-based technologies have begun to tackle this scaling issue. For example, there are a number of automatic grading systems for programming assignments [56]. However, there are limited automatic assessment approaches for open-ended work in other domains, and there is usually no natural language feedback offered to students [8, 75]. In addition, a recent review paper suggests several major limitations of existing automatic question generation techniques for educational purposes [8, 75], including that 1) they target lower cognitive skills, such as fact questions and fill-in-the-gap questions, etc. 2) Meaningful feedback is often missing in these existing systems. 3) Most techniques are domain-specific. Intelligent tutoring systems can adaptively select problems for students and offer feedback, but they also require huge authoring efforts upfront [5, 7]. We see that prior instructional approaches that rely on human efforts are hard to scale and that prior AI-based approaches do not provide high-quality learning experiences. In my dissertation work, I explore techniques that enable the creation of learning opportunities that achieve scale and quality at the same time, by leveraging the complementary strengths of humans and machine intelligence.

This dissertation is organized as follows. In Chapter 3, Chapter 4, Chapter 5, Chapter 6, I present the insights we have gained over the past few years in designing, developing and testing a technique that supports the creation of quality learning opportunities at scale, **through using student solution examples to semi-automatically generate authentic multiple-choice questions for deliberate practice of higher-order thinking**. In Chapter 7, I present theorized analyses and empirical findings regarding the instructional choice of when to use evaluation-type exercises to support student learning, which can be produced at scale using the technique introduced in this dissertation. In Chapter 8, Chapter 9, I present two applications using the insights we have acquired in two other domains.

To support the creation of high-quality learning opportunities, we must first understand what “high quality” means and what are instructors’ current practice in authoring learning materials. The “high-quality” feature I focus on in my dissertation is providing deliberate practice opportunities with immediate feedback to students. A large body of literature suggests focused practice targeting

specific skills helps novices become experts[9, 34]. Prior work also suggests that appropriate scaffolding [132], timely feedback [42, 70, 74], and active engagement [22, 72] are helpful for learning. However, we also observe that in colleges, open-ended assignments are widely used and are treated as a major source of practice and learning in many courses (Chapter 3, Chapter 5, Chapter 7). Considering a regular course expected to take a student 12 hours each week. Besides the time to attend lectures and complete readings, students are usually expected to spend at least 6-9 hours each week on their assignments. This is about 50% - 75% of students' learning time. Through a survey with 22 HCI professors, we found that open-ended assignments are often given and preferred (Chapter 7). **However, timely feedback is one feature of deliberate practice missing in almost all open-ended assignments.** From the students' perspective, open-ended assignments are often large and require multi-steps. Because they require manual grading, students often receive feedback days later rather than immediately after. This discrepancy shows that the learning opportunities we are offering students do not align very well with what theories have suggested.

From the teachers' perspective, when they deploy what they believe are high-quality, authentic, open-ended activities in their courses, a substantial workload is required to grade student solutions and offer feedback to them. Probing into the reasons behind this through interview and surveys with instructors (Chapter 5, Chapter 7), we find that **instructors find the process for writing questions that offer focused practice and immediate feedback to be time-consuming.** We also find that **instructors believe open-ended assignments are better for student learning, which is contradicted by student performance data** (Chapter 7). This also suggests that experts have blind spots that may prevent them from seeing what is best for novices (students).

These discrepancies point to opportunities that if we make creating deliberate learning opportunities easier and more manageable, it may increase the adoption of better pedagogies in college instruction. Towards this goal, I explore methods that source authentic student open-ended solutions to create learning opportunities that offer deliberate practice and immediate feedback and also target higher-order thinking for students. In this dissertation, I present insights from the iterative design, development and testing process of a technique UpGrade, during 3 years engaging with 10 instructors and 600+ students at CMU. UpGrade uses student solution examples to semi-automatically generate authentic multiple-choice questions for the deliberate practice of higher-order thinking. The idea of UpGrade originated when I was a teaching assistant for a research methods course. When I was grading students' assignments, I realized that different teams often had the same mistakes and I was mainly copying and pasting the same piece of feedback to a lot of students. The same experience happened again during my second teaching assistantship. I then conducted research to provide a better formative assessment experience for students, which also saves repetitive efforts from instructors.

Before I developed UpGrade, I conducted a formative interview study (Chapter 5) with instructors and learning engineers to understand their current practices and challenges with designing assessments. The major takeaways include that open-ended projects can be overly challenging, and thus are not good learning opportunities for students. Students show similar mistakes, and mistakes repeatedly appear within one student. When writing assessments, instructors find coming up with scenarios and creating examples to be hard. The design consideration for the system is to source examples from student solutions. Since we are anticipating common student errors, giving them practice opportunities beforehand to avoid these errors can be powerful, and UpGrade

Workflow of UpGrade

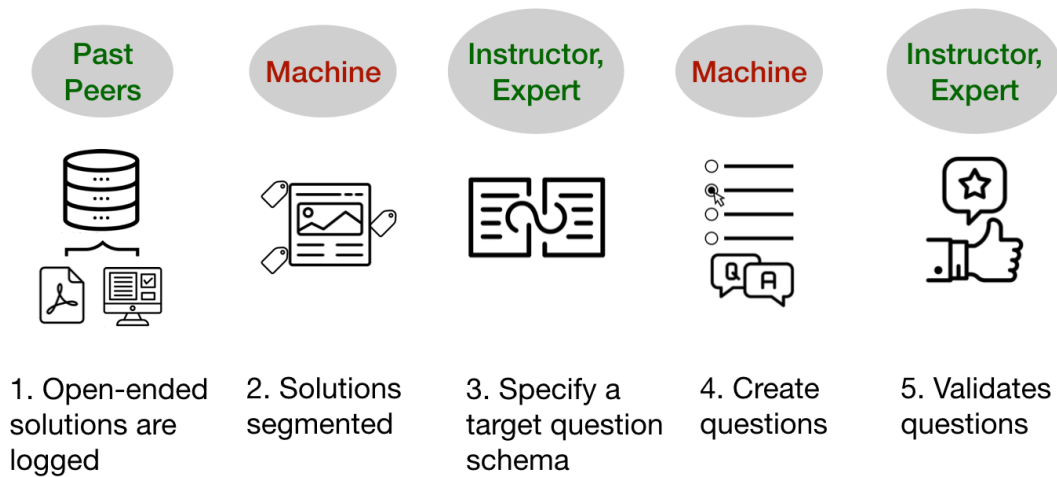


Figure 1.1: Five Step Workflow of UpGrade. Step 1 almost happens naturally. Past students' written solutions are logged, they can be logged in PDF formats or through online forms. In step 2, the system segments the written solutions into components based on the assignment template (for example, through keyword extraction). This step requires a manual check to make sure the data is meaningfully segmented. In Step 3, we ask the instructor to specify a question creation schema, in which the instructor will specify which components in the source will be used in the target question. In step 4, the system reorganizes the source data based on the schema and leverages Natural Language Processing (NLP) techniques to create multiple-choice questions. In the last step, we ask instructors to select questions from the large question pool.

naturally identifies common errors by drawing on past student solutions and reuse instructor feedback.

Fig 5.1 conceptually summarizes the workflow of UpGrade, UpGrade is leveraging the capabilities of both humans and the machine. Past students function as a crowd that offers data sources. Machine segments, selects, and reorganizes examples. The expert validates the questions, in the end, as a quick quality control. This workflow enables UpGrade to quickly produce good quality multiple-choice questions at scale. On the one hand, UpGrade provides deliberate practice for students towards mastery learning with appropriate scaffolding, real-time feedback and repeated practice opportunities. On the other hand, UpGrade aims at reducing repetitive effort from instructors and complementing instructor expertise and effort with machine and crowd intelligence.

UpGrade was first presented in [139] (**Chapter 3**) as a technique that requires a combination of offline (manual) and online (automatic) efforts to produce a large pool of multiple-choice questions. We did two evaluation studies to examine the learning benefits of UpGrade-produced multiple-choice questions for students. The first experiment is presented in **Chapter 3**. In this study, we found that students learnt the same from UpGrade as a traditional open-ended assignment in 30% less time. The learning outcome is measured by a quiz composed of both multiple-choice and open-ended questions. We were also interested in whether the skills exercised through

View Assignment
Create Question Group
Create Question Group (A)
Review Question
Final Question Pool

Storyboard

Answer 1

Storyboard 1: A three-panel comic strip. Panel 1: A woman on a train looking at a phone. Panel 2: A woman on a train looking at a phone. Panel 3: A woman on a train looking at a phone.

User Need for the Storyboard

Answer 1

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Answer 2

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Answer 3

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Answer 4

Riders need a reliable centralized channel with all shuttle information

Answer 5

Lead Question for the Storyboard

Target Question Stem:

- Textbox | - Select Component

Which of the following user need is this storyboard designed for?

C

None

Storyboard

User Need for the Storyboard

Lead Question for the Storyboard

Target Correct Answer:

- Select Component

None

Storyboard

User Need for the Storyboard

Lead Question for the Storyboard

Target Incorrect Answer:

- Select Component | - Criteria | - Number

None

Storyboard

User Need for the Storyboard

Lead Question for the Storyboard

None

Random

Similar

Different

3

Which of the following user need is this storyboard designed for?

1

Storyboard 1: A three-panel comic strip. Panel 1: A woman on a train looking at a phone. Panel 2: A woman on a train looking at a phone. Panel 3: A woman on a train looking at a phone.

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Riders want to know the waiting time for their shuttle in order to determine which transportation method to use. Similarity:0.6123724356957946

Riders need to know the current capacity of the buses to make better decisions. Similarity:0.3849001794597506

Having more confidence in how to ride the shuttle when boarding Similarity:0.3077287274483319

Figure 1.2: QuizMaker Interface: Segmented past student work is displayed to the left; Instructors can create question schemata and preview the question on the right.

evaluation-type activities could transfer to improved task performance. We then ran a second evaluation experiment investigating whether exercising with UpGrade could lead to improvements in the quality of authentic and complex open-ended work. The result for the second experiment is presented in **Chapter 4**. We found that exercising with UpGrade before doing the open-ended assignment (creating Storyboards for speed dating studies) helped improve the quality of open-ended work students produced, i.e., higher quality of storyboards and protocols for subsequent speed dating studies. These studies demonstrate the learning benefits of UpGrade-produced practice questions. And at the same time, following the workflow of UpGrade, instructors can create hundreds of these questions very quickly.

In **Chapter 5**, I describe the iterative design process of the authoring interface of UpGrade (as shown in Fig 5.3) and present the takeaways around using UpGrade in practice with instructors and learning engineers. UpGrade works by taking advantage of the complementary strengths of peers, machine and instructors, and the content creation process wouldn't have been possible without each stakeholder here. With instructors alone, it is often hard to write lots of elaborated good examples and wrong answers. With machine alone, we see that existing question generation

Human-Machine Partnership for Content Creation

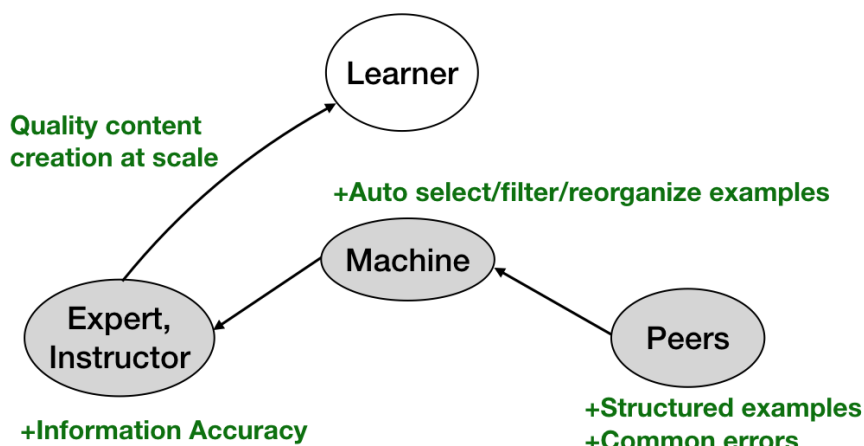


Figure 1.3: UpGrade takes advantages of the strengths of peers, machine and instructors.

techniques only produce fact questions and are not flexible. Here, past peers offer a powerful data source with structured examples that show common errors. With this better data input, machine can auto select/filter and reorganize examples. Instructors, in the end, will check information accuracy and comprehensiveness. This altogether supports quality content creation at scale, as shown in Fig 1.3.

To enhance quality control of the questions created by UpGrade, we investigated psychometric approaches using student performance data to automatically prune out low-quality questions. In study 3 (**Chapter 6**), we collected learner responses to UpGrade-created questions, and used Cronbach's alpha [27] to identify unreliable question items. After pruning out low-quality questions, UpGrade produces a question bank that exceeds reliability standards for classroom use. We demonstrate that crowd (such as MTurk) can be leveraged as a source for quality control. We also demonstrate that when reducing the sample size, the number of good items predicted as bad increases, but the number of bad items predicted as good remains the same. This suggests that with techniques such as UpGrade, which produces a large question pool, the system can use relatively small data sets to prune out unreliable items without worrying about having false positives.

As I explore the feasibility of applying UpGrade to a variety of courses, there arises the question of when are multiple-choice practice questions appropriate? In study 4 (**Chapter 7**)[140], I present theorized analyses and empirical findings and suggest when to use evaluation-type exercises such as multiple-choice questions from a theoretical angle. On the one hand, we see encouraging results suggesting multiple-choice questions can be useful. On the other hand, we surface a negative sentiment towards multiple-choice questions from instructors, e.g., instructors tend to believe "Although open-ended assignments are less scalable, they provide better learning opportunities compared to multiple-choice assignments." We conducted a series of studies in four courses investigating when would multiple-choice questions be appropriate and applicable following theorized analyses of the different types of cognitive efforts required when answering

questions. We also contrasted expert prediction on the difficulty of these practice questions with student performance data. Our studies demonstrate that “evaluation” efforts are critical in at least some problem-solving processes, and in-fact in all the problem-solving processes we have investigated. We suggest that for at least these “evaluation-heavy” domains, online learning can benefit from the scaling advantages of evaluation-type exercises, e.g., multiple-choice questions, without sacrificing (and perhaps gaining) learning quality. We also observe that instructor prediction of question difficulty doesn’t align with student performance. This is another case that suggests expert blind spots exist. I demonstrate that well-designed, low-effort experimental comparisons can help experts (instructors) make more accurate and nuanced instructional decisions and designs.

In **Chapter 8**, I present the design of a teacher training system [141] using insights on exercising evaluative cognitive efforts found in study 4 (Chapter 7). With my studies in Chapter 7 and other prior work, we have shown that evaluating the quality of solutions can support learning and performance on generating solutions afterwards, even with higher learning efficiency compared with practicing with generating solutions only. For example, Yannier *et al.* shows that evaluating “which towers would be 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 [152]. Ericsson *et al.* 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 prior work is mostly focused on technical skills that do not require interpersonal communication. For skills such as asking questions, it remains unknown whether evaluating responses can be a useful exercise, and whether it is more, less, or equally useful for learning as generating improvisational responses to scenarios. In Chapter 8, I present the design of a system that helps teachers learn questioning strategies and I present findings regarding having teachers evaluate transcripts achieved similar benefits as asking teachers to generate questions in chats.

In **Chapter 9**, I present a second example showing that with close human and machine collaboration, we can repurpose online videos as software tutorials for end-users [138]. Besides formal higher education settings, learning is also happening ubiquitously as people watch videos, read online articles, etc. In Chapter 7, I describe a technique where I harness examples from online videos and logs to support the use of complex graphical software. This workflow categorization and recommendation technique also takes advantage of the complementary strengths of peers, experts and machine intelligence, as shown in Fig 1.4.

Human-Machine Partnership for Content Creation

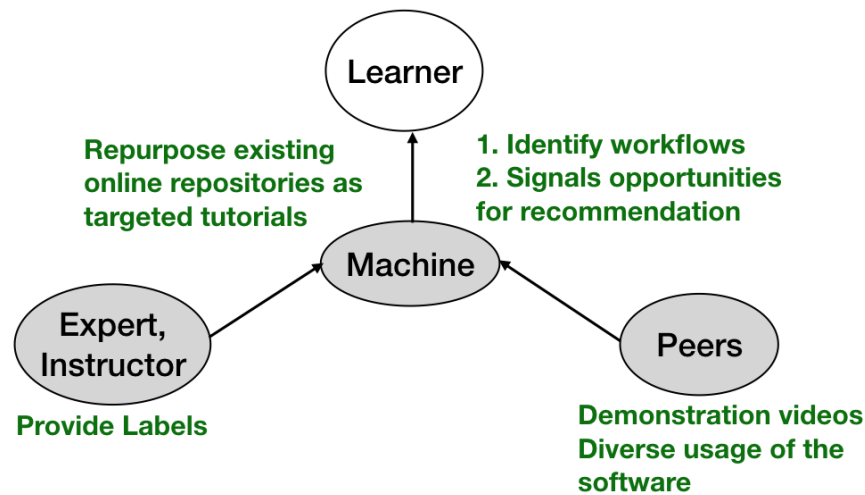


Figure 1.4: In this case, peers who can be users of the software everywhere in the world contribute demonstration videos. We collect these videos (and their command logs) as input and use machine to identify an end user’s workflow, expert input is elicited in this process. With this approach, existing online repositories may be repurposed as targeted tutorials for end users

I summarize the contributions this dissertation makes below. I will elaborate on the suggestions and related future work in Chapter 10 and Chapter 11.

- This dissertation contributes a novel technique that uses student solution examples to semi-automatically generate authentic multiple-choice questions for deliberate practice of higher-order thinking. With two classroom experiments, we show that the deliberate practice opportunities created with this technique help students gain conceptual understanding more efficiently and help improve the quality of student open-ended work.
- This dissertation contributes insights about developing effective learning at scale systems by leveraging the complementary strengths from peers, experts, and machine intelligence, differentiating it from existing systems that solely rely on machine or crowds. There are three components that contribute to the effectiveness of this learnersourcing technique. First, instructors are not good at creating distractors, and actual elaborated student errors are helpful as sources. Second, we apply simple natural language processing techniques to select distractors of interest. Third, we involve instructors closely in the process to enhance quality.
- Applying the workflow of UpGrade in practice across courses demonstrates the generalizability and practical value of this approach and helps inform the design of an interface to facilitate the independent use of UpGrade by instructors for authoring practice questions.
- When instructor efforts are not available for reviewing and revising the questions to enhance quality control, we demonstrate an effective quality control method using psychometric approaches to automatically select high-quality question items from a large question pool.

- I demonstrate two examples where complementary human and machine intelligence are leveraged to create educational materials at scale. In both cases, crowds (e.g., past students) offer a powerful data source with structure examples that show common errors. Machine automatically selects, filters and reorganizes examples. Experts (instructors) check information accuracy and comprehensiveness to enhance content quality.
- This dissertation suggests another layer to further distinguish knowledge components, by the required generation and evaluation efforts in problem-solving. The practical implication for a more nuanced understanding of knowledge components (KCs) is to help instructors make more nuanced and accurate instructional decisions, e.g., using “evaluation-type” exercises for evaluation-heavy skills.
- This dissertation indicates that, at least for some domains, 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.
- This dissertation provides further evidence that instructors have so-called “expert blind spots”, revealed through cases where their beliefs and student performance do not match. Specifically, instructors believe open-ended assignments to be better for student learning, which is contradicted by student performance data. 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.
- This dissertation also makes suggestions to the model of take-home assignments used in a higher-education context, especially relevant to topics similar to the ones we have investigated. We surface an issue that open-ended work students turn in are of low quality, suggesting there are cases when students are not ready and need scaffolding before they do complex open-ended work. An alternative model would be giving students deliberate practice opportunities before the assignment of flexible open-ended work. The deliberate practice opportunities can be easily created with techniques such as UpGrade.
- I present the design of a system that benefits from the distinction between generation and evaluation efforts during problem-solving. ELK (Eliciting Learner Knowledge), is a text-based role-playing system that enables pre-service teachers to practice questioning moves through simulated “teacher-student” conversations.

Chapter 2

Related Work

In this chapter, I review relevant literature that my work situates in. First, I review literature in recent advances in automatic question generation techniques for educational purpose. Second, I review literature in the emerging field of learnersourcing [60], including systems that have been developed by sourcing existing student data [63, 64, 65, 89, 143, 146]. Third, I review literature related to skill acquisition, more specifically, theory of deliberate practice [33, 34, 35] and constraint-based expertise acquisition [68, 102, 103]. Fourth, I introduce the Knowledge-Learning-Instruction framework [70], which I followed to describe the relationships between knowledge, learning and instruction. Fifth, I review literature related to learning problem-solving skills [100], more specifically the use of worked examples [9, 21, 104, 129] and the role of feedback [9, 51, 67, 73]. Finally, I review psychometric approaches [27, 50] which we applied to enhance quality control in UpGrade.

2.1 Automatic Question Generation for Educational Purposes

This section discusses prior work on automatic generation techniques for educational purposes. The discussion is based on two recent literature review papers on the topic, one was published on 2015, and the other was published early this year summarizing the advances in the area since 2014. [8, 75]. Some of the major findings include that 1) existing automatic question generation techniques produce questions that target lower cognitive skills, such as fact questions and fill-in-the-gap questions, etc. 2) Meaningful feedback is often missing in these existing systems. 3) Most techniques are domain-specific, and there are more techniques focusing on language learning, followed by math and medicine when there are existing knowledge bases that provide language ontology through NLP tools to support question creation. 4) The main purpose of generating questions is to use them as assessment, for example, in exams. Few projects used the generated questions as a way of instruction.

Prior work mostly focused on domain-specific question generation techniques, especially on language learning and medicine learning. Generating questions for a specific domain is more prevalent than generating domain-unspecific questions. For example, there are many techniques/studies for generating language learning questions, followed by math and medicine. For language learning, there are standardized tests developed by professional organizations such

as ETS. These techniques use NLP tools for shallow understanding of text with an acceptable performance, e.g., changing the verb form of the key, “write”, “written”, “wrote” as distractors for “writing.” Another plausible reason for interest in questions on medicine is the availability of NLP tools (named entity recogniser and co-reference resolvers) for processing medical text. External, structured knowledge sources are needed to find what is true and what is similar. In terms of question types, simple factual wh-questions and gap-fill questions are the most generated types of questions. The types of questions generated from ontologies are more varied than the types of questions generated from text.

There are several limitations in prior work identified, 1) there is limited research on controlling the difficulty of generated questions, and on generating informative feedback; 2) the quality of the questions generated has been a concern; 3) the simplicity of generated questions is another concern, which has also been highlighted in [124]. Most generated questions consist of a few terms and target lower cognitive levels. While these questions are still useful, there are potentials for improvement by exploring the generation of other, higher order and more complex, types of questions.

The strategy these approaches primarily use is limited in that they tend to simply transform given text from declarative statements to questions. In this work, we explore data inputs that are not declarative statements, but elaborated solutions from students that display common misconceptions with accompanying thought processes. On the other hand, we involve experts at multiple stages. Instructors specify question creation schema that target higher order thinking, e.g., evaluation. Instructors also review questions and provide feedback in the end. With the introduction of these two new components, our technique produces higher quality content.

2.2 Learnersourcing

The idea of learnersourcing, proposed and implemented in [60], is a form of crowdsourcing in which learners collectively contribute novel content for future learners while engaging in a meaningful learning experience themselves. For example, LectureScape [63] helps learners navigate online lecture videos using interaction data aggregated over all previous video watchers. ConceptScape [89] generates and presents a concept map for lecture videos through prompting video watchers to externalize reflections on the video. AXIS [146] asks learners to generate, revise and evaluate explanations as they solve a problem, and then presents these explanations to future learners. Other crowdsourcing workflows are designed to extract step-by-step information [64] from how-to-videos or construct subgoals [143] to enhance existing how-to videos.

Prior work used learnersourcing to enhance video watching experience and offer explanations to students. A gap in this literature that our work seeks to fill is that students’ written assignments have not been explored yet as a source for benefiting future learners. Written assignments often take hours of student time to complete, containing rich information, thus could be used a valuable input for learnersourcing.

2.3 Expertise Development

2.3.1 Deliberate Practice

The theory of deliberate practice suggests that a sufficient amount of experience or practice may not lead to maximal performance, such as merely executing proficiently during routine work. Instead, further improvements on expertise depend on deliberate efforts to change particular aspects of performance, in other words, practice that involves elements of the desired competence that are at the edge of learners' capability is most valuable for expertise development. [34, 35] The theory of deliberate practice suggests that to support learners in expertise development, learning opportunities should be created that would involve elements of the desired competence that are at the edge of learners' capability. Deliberate practice also involves the provision of immediate feedback, time for problem-solving and evaluation, and opportunities for repeated performance to refine behavior.[33] However, instructional events, such as open-ended assignments, do not support learning events with deliberate practice.

2.3.2 Constraint-based Expertise Acquisition

Constraint-based expertise acquisition theory explains knowledge as a sets of constraints. It suggests learning to solve a problem can be viewed as learning to meet and avoid certain constraints. For example, opening a door is normally subject to the constraint that the door should not become damaged in the process. [68] Faced with an unfamiliar problem, the problem solver knows neither what to do nor what to avoid. If the problem reminds them of some problem encountered in the past, constraints relevant to that problem are likely to be activated. If the constraints can be adapted to the unfamiliar problem, the active constraints might circumscribe a problem space, in which the problem solvers should search for a solution. Following the theory of constraint-based expertise acquisition, when novices are learning new problem-solving skills, leveraging examples that meet or avoid certain constraints as learning materials may help them learn the constraints well.

Constraint-based Student Modeling

The student modeling problem in its general form can be stated as follows [102]: Given a behavioral record (of some sort), infer what the student knows and does not know about the relevant topic. [103] Ohlsson proposed constraint-based student modeling in contrast to other student modeling techniques, e.g., model tracing techniques [10]. The advantage over model tracing techniques is that it can make inference about student knowledge as the constraint is violated.

The main goal for student modeling is to guide subsequent pedagogical decision making. In this sense, the system may not need to differentiate between some mistakes if they point to the same instructional approach. [103] This implies that a student modeling approach where student are modeled in terms of equivalence classes of solutions rather than specific solutions or strategies. Constraint-based student modeling is based on the notion that the student can be described in terms of entities more abstract than particular solution paths or strategies.[103] Based

on these considerations, constraint-based student modeling is proposed. Each constraint defines two equivalence classes: solutions that violate the constraint and solutions which do not. If the constraints are chosen to represent fundamental ideas of the domain, then these two classes of situations require different tutorial responses.

For example, when a student is working on the problem of $1/4 + 2/3 =$. If the system observes student put $3/$ on the right of the equation, that suggests thte student is simply adding the numerators. It violates the constraint that if two fractions have unequal denominators, then the denominators must be equalized before the fractions can be added. Violating this constraint indicates that the student is not thinking about the denominator as the unit in terms of which the numerator is expressed.

We consider many evaluation-heavy content domains can benefit from constraint-based student modeling. For example, when students are learning to design survey questions. They are essentially learning the constraints associated with survey question design, e.g., a survey question shouldn't be a leading question, shouldn't ask respondents to estimate, shouldn't ask a double-barreled question, etc. To model student knowledge in terms of which constraints are violated would be more effective than modeling how students came up with a survey question, because the solution space is infinite. In my thesis, I apply the theory of constraint-based expertise acquisition to support learners in learning evaluation-heavy skills.

2.4 The Knowledge-Learning-Instruction Framework

The Knowledge-Learning-Instruction framework is a widely adopted framework that specifies three taxonomies, kinds of knowledge, kinds of learning processes, and kinds of instructional choices, and dependencies between them. The framework demonstrates how kinds of knowledge constrain learning processes and how these processes constrain which instructional choices will be optimal in producing robust student learning.[70] It is a widely adopted framework to generate and test research questions within specific domains and instructional situations. In my thesis, I follow the definition of instructional event, assessment event, learning event, and knowledge components, as defined in the widely adopted Knowledge-Learning-Instruction framework [70]. More specifically, instructional events and assessment events are designed and delivered by instructors, e.g., a lecture is an instructional event, and an exam is an assessment event. Learning events refer to the learning processes students engage in behind scenes, which are not observable. Students gain knowledge components (KCs) through learning events, which can be inferred from performance on assessment events. The KLI framework suggests that kinds of KCs drive instructional event choices. For example, instructional approaches that emphasize recall and spacing of practice may benefit learning of historical facts, vocabulary; whereas instructional approaches that prompt self-explanation in students would be more valuable for learning complex principles, such as Newton's laws.

2.5 Problem Solving, Worked Examples, Feedback

Newell and Simon [100] proposed a "generate and test" problem-solving approach, which suggests that when people solve problems they carry out selective search in a problem space that incorporates some of the structural information of the task environment. They propose that the set of human behaviors of "problem solving" encompasses both the activities required to construct a problem space in the face of a new task environment, and the activities required to solve a particular problem in some problem space, new or old. The proposition that "search in a well-defined problem space is not problem solving at all" was found to be empirically false. This aligns with the distinction between generation and evaluation processes we are making in problem-solving. In particular, we found domains where learning to evaluate alternative solution options is much harder than learning to generate candidate solutions, and thus practicing the "evaluation" aspect should be emphasized in learning.

The theory of worked examples suggests that when a problem to be solved is sufficiently demanding, students may not have enough cognitive resources to learn from solving the problem [9, 104]. Providing instructional scaffolding to a practice activity promotes learning when it helps students practice the target skills at an appropriate level of challenge [21]. Worked examples [129] are one such type of scaffold, which frees up cognitive resources and allows students to see the key features of a problem and analyze the steps and reasons behind problem-solving. In my work, I am developing UpGrade that provides such instructional scaffolding through auto-created worked examples.

Targeted feedback is also critical during deliberate practice. Many studies have shown that feedback interventions improve learning more than non-feedback ones [67]. Generally, more frequent feedback leads to more efficient learning because it helps students stay on track [51]. However in practice, crafting deliberate practice opportunities with frequent feedback requires careful design and substantive effort from instructors. Furthermore, for open-ended problems, provision of frequent feedback may not be affordable, especially in large-scale classes [73]. In this work, we design UpGrade to offer deliberate practice on open-ended problems without the need for instructors to put in hours of effort in the preparation or during use. One risk of focused, deliberate practice opportunities is that the focused nature might preclude the experience of authentic activities [9]. UpGrade addresses this concern by delivering deliberate practice that is situated within authentic activities.

2.6 Quality Control through Psychometric Approaches

Prior work has used learner subjective ratings [146] to select high quality content in learner-sourcing systems. In this work, we instead explore psychometric methods to evaluate question reliability using student performance data. Common psychometric methods evaluate test reliability by the internal consistency of question items within a test, e.g., using a Rasch model [148], Item Response Theory (IRT) model [50], or Cronbach's alpha [27]. If question items within a test are consistent in measuring student capabilities or in differentiating knowledgeable and less knowledgeable students, the test is considered reliable and question items are considered to be of high enough quality. On the other hand, if a question item is failing knowledgeable students but

favoring less knowledgeable students, the question item is considered to be problematic and needs redesign. Cronbach's alpha is the most common internal consistency measure, and is incorporated in UpGrade to evaluate the internal consistency of questions generated. An acceptable reliability score (Cronbach's alpha) for exams is in the range of 0.7-0.95 [130]. As reported in the 2011 TOEFL iBT research report [36], the reliability estimate for TOEFL iBT Speaking and Writing sections are 0.84 and 0.8 respectively, measured by Cronbach's alpha. We expect a reliability score in the range of 0.7-0.8 to indicate good enough internal consistency of a test and the question items in the test to be reliable.

Chapter 3

UpGrade: A Learnersourcing Technique

In schools and colleges around the world, open-ended homework assignments are commonly used. However, such assignments require substantial instructor effort for grading, and tend not to support opportunities for repeated practice. We propose *UpGrade*, a novel learnersourcing approach that generates scalable learning opportunities using prior student solutions to open-ended problems. UpGrade creates interactive questions that offer automated and real-time feedback, while enabling repeated practice. In a two-week experiment in a college-level HCI course, students answering UpGrade-created questions instead of traditional open-ended assignments achieved indistinguishable learning outcomes in ~30% less time. Further, no manual grading effort is required. To enhance quality control, UpGrade incorporates a psychometric approach using crowd workers' answers to automatically prune out low quality questions, resulting in a question bank that exceeds reliability standards for classroom use.

3.1 Introduction

A key insight that has spawned a new direction in crowdsourcing research called *learnersourcing* is that learners around the world unwittingly produce content that can be leveraged to create novel learning opportunities. For example, video watching traces [63], video annotations [64, 89, 143], or explanations [146] generated by prior learners were sourced to benefit future learners. In this paper, we explore written homework assignments as a new and powerful input for learnersourcing. Afterall, students are producing great volumes of written content in response to open-ended assignments. We describe how this content can be automatically transformed into online practice activities where student learning is supported through immediate feedback and we present evaluations of the quality of the questions created, the learning outcomes achieved, and time savings for students and instructors.

Open-ended assignments are widely used in schools and colleges as formative assessments. They typically involve qualitative feedback offered by instructors, and are designed to inform subsequent learning in contrast with summative assessments, such as exams. At the same time, open-ended assignments require substantive efforts from instructors to grade and provide feedback. Furthermore, the full benefits of this feedback is best realized when it is provided soon after students complete assignments and when they are given the opportunity to incorporate feedback

into further practice. However, timely return of detailed feedback is hard to achieve and open-ended assignments are often used as a one-off activity whereby there is little or no chance for deliberate practice on concepts or skills that were not demonstrably mastered.

In this work, we propose *UpGrade*, a novel learnersourcing approach that delivers scalable and efficient learning opportunities, reducing time commitment from both students and instructors. Following the workflow of UpGrade, instructors can create hundreds of multiple-choice questions from prior student solutions to open-ended problems with minimal effort. UpGrade-created questions also offer real-time feedback for repeated practice. UpGrade can be used as an alternative or primer to traditional open-ended assignments, with more instructional scaffolding towards mastery of the knowledge and skills. UpGrade works by (i) chunking information to be learned into smaller pieces, which allows novices to gradually engage more information; (ii) enabling deliberate practice, which helps novices to develop mastery on knowledge and skills; and (iii) offering immediate and frequent feedback, which helps students stay on track and addresses their errors as they occur.

To evaluate UpGrade in a realistic learning setting, we applied it in a college-level Human-Computer Interaction (HCI) course that teaches user-centered research methods, of which we focused on heuristic evaluation and survey design. In a two-week classroom experiment using a crossover design, we demonstrated that students answering interactive UpGrade-created multiple-choice questions instead of traditional open-ended assignments achieved indistinguishable learning outcomes, while reducing assignment completion time by ~30% and removing the need for instructor grading. This first classroom experiment of UpGrade demonstrates substantial promise for the approach. We also explore crowdsourced methods for evaluating and enhancing the quality of the automatically generated questions. UpGrade incorporates a psychometric method to distinguish reliable versus unreliable question items. Unreliable question items were successfully identified through a validation study with 70 participants on Amazon Mechanical Turk. This results in a reliable question bank with an internal consistency that exceeds the standards for classroom use.

In summary, we make the following key contributions:

- New technique: UpGrade, a learnersourcing approach that delivers scalable and efficient learning opportunities, reducing time commitment from both students and instructors.
- Evidence of support for learning: An experiment of UpGrade, demonstrating effective time reduction for students and instructors, while achieving indistinguishable learning outcomes compared to traditional open-ended assignments.
- Approach for quality control: An effective quality control method for automatically selecting high quality learning materials with minimal crowdsourcing effort.

3.2 Related Work

Our work extends the frontier of work in an emerging area of crowdsourcing referred to as learnersourcing [60, 63, 64, 89, 143, 146]. The design of UpGrade is motivated by learning theories related to instructional scaffolding [9], worked examples [129], and deliberate practice [34]. To lay a theoretical foundation for our work, in this section we discuss the cognitive

processes involved in solving multiple-choice and open-ended problems. From a more practical standpoint we discuss how frequent feedback and deliberate practice are not always affordable for open-ended problems [73]. To address potential concerns that an automated approach to item generation introduces the risk of unreliable or poor quality items, we reviewed established psychometric methods to evaluate test reliability, which informs our quality control approach.

3.2.1 Learnersourcing Techniques

The idea of learnersourcing, proposed and implemented in [60], is a form of crowdsourcing in which learners collectively contribute novel content for future learners while engaging in a meaningful learning experience themselves. For example, LectureScape [63] helps learners navigate online lecture videos using interaction data aggregated over all previous video watchers. ConceptScape [89] generates and presents a concept map for lecture videos through prompting video watchers to externalize reflections on the video. AXIS [146] asks learners to generate, revise and evaluate explanations as they solve a problem, and then presents these explanations to future learners. Other crowdsourcing workflows are designed to extract step-by-step information [64] from how-to-videos or construct subgoals [143] to enhance existing how-to videos.

Prior work used learnersourcing to enhance video watching experience and offer explanations to students. A gap in this literature that our work seeks to fill is that students' written assignments have not been explored yet as a source for benefiting future learners. Written assignments often take hours of student time to complete, containing rich information, thus could be used a valuable input for learnersourcing.

3.2.2 Worked Examples and Scaffolding

UpGrade addresses two important issues related to design of effective scaffolding, one related to cognitive load and the other related to expert blind spots. First, though open-ended work provides opportunities for authentic learning experiences, a downside is that these rich experiences may consume most of a student's available cognitive load when they have not mastered the skills and knowledge needed to be successful at the activity [9]. If the problem itself is sufficiently demanding, students may not have enough cognitive resources to learn from solving the problem [104]. Providing instructional scaffolding to a practice activity promotes learning when it helps students practice the target skills at an appropriate level of challenge [21]. Worked examples [129] are one such type of scaffold, which frees up cognitive resources and allows students to see the key features of a problem and analyze the steps and reasons behind problem-solving. UpGrade provides instructional scaffolding in support of open-ended problem-solving through auto-generated worked examples.

A second concern is expert blind spots [118], where the teachers' expertise makes it difficult for them to anticipate the specific needs of their students. This may prevent instructors from authoring scaffolded learning experiences that take into account all the component skills and knowledge required for complex tasks. On the other hand, prior solutions might provide a complementary source of insight, offering visibility into common mistakes and misconceptions. This motivates the design of UpGrade to decompose student solutions and display the merits or mistakes of the solutions for future students' reference.

3.2.3 Repeated Practice and Feedback

Deliberate practice, which is focused practice targeting specific skills, assists novices in becoming experts [118]. Research shows that the amount of time a learner spends in deliberate practice rather than more generic practice is what predicts continued learning in a given field [34]. By breaking information down into bite-sized chunks, deliberate practice allows novice learners to gradually engage more information without being overwhelmed [118]. Targeted feedback is also critical during deliberate practice. Many studies have shown that feedback interventions improve learning more than non-feedback ones [67]. Generally, more frequent feedback leads to more efficient learning because it helps students stay on track [51].

However in practice, crafting deliberate practice opportunities with frequent feedback requires careful design and substantive effort from instructors. Furthermore, for open-ended problems, provision of frequent feedback may not be affordable, especially in large-scale classes [73]. In this work, we design UpGrade to offer deliberate practice on open-ended problems without the need for instructors to put in hours of effort in the preparation or during use. One risk of focused, deliberate practice opportunities is that the focused nature might preclude the experience of authentic activities [9]. UpGrade addresses this concern by delivering deliberate practice that is situated within authentic activities.

3.2.4 Quality Control Methods

Prior work has used learner subjective ratings [146] to select high quality content in learner-sourcing systems. In this work, we instead explore psychometric methods to evaluate question reliability using student performance data. Common psychometric methods evaluate test reliability by the internal consistency of question items within a test, *e.g.*, using a Rasch model [148], Item Response Theory (IRT) model [50], or Cronbach's alpha [27]. If question items within a test are consistent in measuring student capabilities or in differentiating knowledgeable and less knowledgeable students, the test is considered reliable and question items are considered to be of high enough quality. On the other hand, if a question item is failing knowledgeable students but favoring less knowledgeable students, the question item is considered to be problematic and needs redesign. Cronbach's alpha is the most common internal consistency measure, and is incorporated in UpGrade to evaluate the internal consistency of questions generated. An acceptable reliability score (Cronbach's alpha) for exams is in the range of 0.7-0.95 [130]. As reported in the 2011 TOEFL iBT research report [36], the reliability estimate for TOEFL iBT Speaking and Writing sections are 0.84 and 0.8 respectively, measured by Cronbach's alpha. We expect a reliability score in the range of 0.7-0.8 to indicate good enough internal consistency of a test and the question items in the test to be reliable.

3.3 Formative Study: Assignment Survey

We first conducted a formative study to understand what commonly-used open-ended assignments look like, and to identify potential cases where sourcing existing examples could be beneficial. We did a content analysis of the assignments of six courses offered to both undergraduate and

graduate students in the Human-Computer Interaction (HCI) program at an R1 institution. We used a qualitative approach to examine the learning goals of these assignments and grouped them into several clusters. We identified cases where the skills to be learned in these assignments could be taught through evaluating examples, as shown in Figure 3.1. We illustrate how we construct the graph below.

The courses we surveyed include two user experience (UX) method courses, two technical (computer science-related) courses, one design course and one learning sciences course. We took a bottom-up approach, mapping out the learning goals and requirements in the assignments. Three clusters of assignments emerged, *(i)* Solve a problem or generate a solution, which most assignments fall in; *(ii)* Learn to use a tool, *e.g.*, get familiar with a software, set up a mobile data collection module; and *(iii)* Share reading reflections and opinions. *(ii)* and *(iii)* were less frequent in the sample and were often not graded, here we focus on the main cluster *(i)*.

Problem-solving assignments include both group projects and individual projects. Individual projects are usually intended for skill building, whereas group projects are for practicing skill integration and content generation. For group projects that involve skill building modules, they assemble that of individual projects. Here we only discuss the branch of individual projects. We saw two types of individual problem-solving projects emerging from the data, open-ended problem solving, asking students to generate a solution to a given problem; and doubly open-ended problem solving, asking students to first define a problem and then generate a solution. We highlight the distinction here because they offer different sources for UpGrade to create multiple-choice questions. Among the problem-solving tasks, some have a single success path or a limited set of success paths, *e.g.*, computing the probability of an event using the Naive Bayes model; computing the mean of a variable in a given dataset. Most problems in our surveyed domains (*i.e.*, UX methods, design, learning sciences) do not have a single success path. This also applies to authentic problem-solving tasks in workplaces.

Traditional computer-based tutors such as Assisments [52] and example tracing tutors [6] were designed for problems with single or limited success paths. UpGrade mainly targets at problems that do not have a single success path. In such problem-solving tasks, students often need to evaluate the solutions they came up with, rationalize why they made the decisions, and revise their solution based on certain criteria. For some domains, the real challenge in solving a problem is to evaluate the quality of a proposed solution rather than to come up with an initial solution. Shifting the practice focus from generating solutions to evaluating existing solutions could be beneficial for learning such skills. We consider such “evaluation-heavy” problem-solving skills (Figure 3.1) could be exercised well through multiple-choice tasks that emphasize evaluation. We listed example skills that are considered to be “evaluation-heavy” in our survey in Table 3.1.

3.4 UpGrade

In this section, we describe UpGrade’s workflow for creating multiple-choice questions from prior students’ open-ended solutions. An overview of the workflow is shown in Figure 3.2. We illustrate each step using an example to offer a proof of concept that this technique can be applied in practice. The example course we used is an HCI research methods course that has been offered in the department for 5+ years. We refer to the course as UX101 for the rest of the paper. We

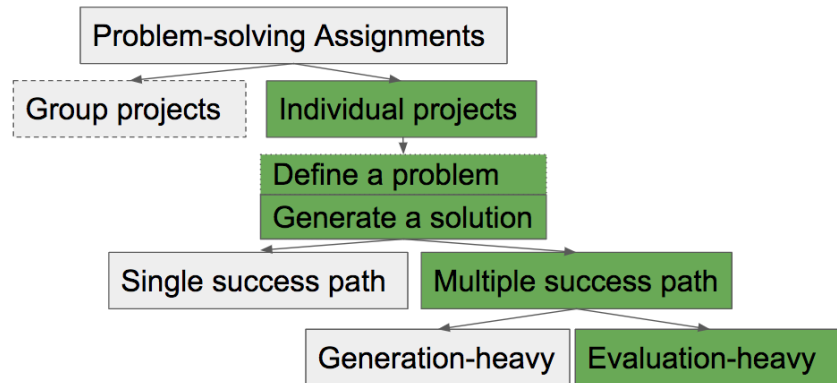


Figure 3.1: Problem-solving assignment classification from 6 HCI courses.

Course type	“Evaluation-heavy” skill
Technical	Propose new features to a model based on error analysis
Learning sciences	Perform a theoretical cognitive task analysis
Design	Ideate concept maps and conceptual models
UX method	Design a survey
UX method	Heuristic evaluation (critique an interface and come up with redesigns)

Table 3.1: Examples of “evaluation-heavy” problem-solving skills.

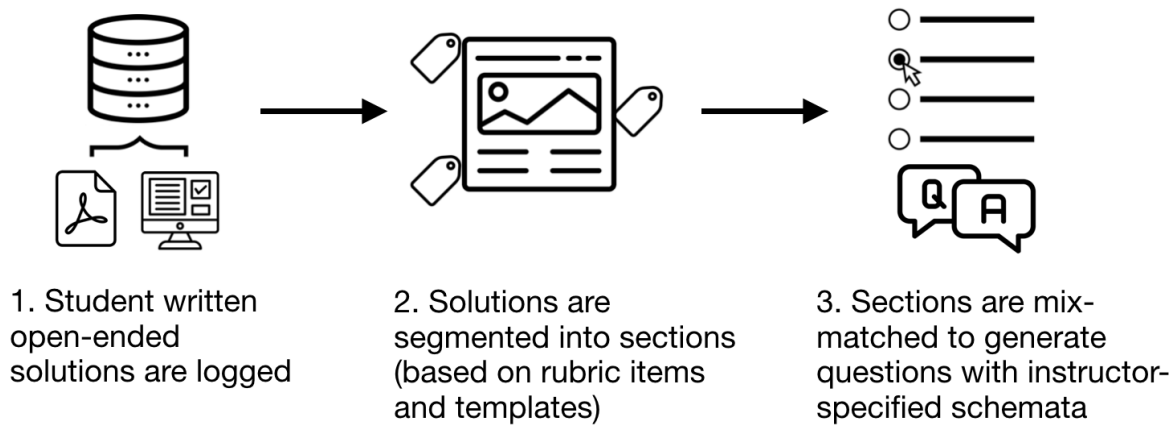


Figure 3.2: UpGrade’s workflow.

focused on two topics of UX101 to create questions, Survey Design and Heuristic Evaluation. Both are “evaluation-heavy” problem-solving skills as categorized in the formative study. Prior offerings of UX101 used one open-ended assignment per topic to help students learn the method. For Survey Design, students were asked to design a survey; for Heuristic Evaluation, students were asked to write a report documenting heuristic problems found for a given website. Past assignment submissions were assessed based on the rubric items shown in Figure 3.3.

3.4.1 Solution Logging

UpGrade requires structured data of students’ open-ended solutions, which can be logged in different formats. We collected all student assignment solutions under the topics of Survey Design and Heuristic Evaluation that were submitted in the 2015 offering of UX101, with ~100 written assignment solutions per topic. All files were in PDF format, the majority of which had a length of 10+ pages, which is typical for college-level open-ended assignments. The assignment solutions were graded and offered feedback to by peers and TAs through an online platform Coursemark based on the assignment rubric (Figure 3.3). Feedback data from Coursemark was scraped in association with rubric item for all the solutions. For courses where students’ open-ended solutions are logged in online forms, the next step for solution segmentation will not be necessary.

3.4.2 Solution Segmentation Based on Assignment Rubric

UpGrade then assigns structures to assignment PDF documents by segmenting the solution based on rubric items. For our collected PDF assignment solutions, UpGrade first converts them to HTML files using the Adobe Acrobat API. UpGrade then employs a Python script to segment the HTML files into sections based on DOM tags and text styles (*e.g.*, <h1>, <h2>, <p>). We found this method to be more effective in this segmentation task than using headings or texts. Different students may use different language to describe each section. However, when they start a new section, the DOM tag or text style is always different from the previous section. Moreover, the segmentation technique also associates in-text images with sections, since image DOM tags (*e.g.*,

) are inside <p> tags. Each assignment solution file is reorganized into a .txt file with one section per line.

The Survey Design and Heuristic Evaluation assignments followed templates. For example, in the Heuristic Evaluation assignment, students were asked to identify five heuristic problems in a given website. For each heuristic problem, it will be evaluated based on five rubric items, including *Description* of the problem, heuristic rule *Violation*, *Explanation* of why the rule is violated, justification of the *Severity* of the problem, and a *Remedy* plan to fix the problem. For solutions whose segmented results matched the rubric items in the template, the segmented sections were automatically associated with each rubric item. However, for solutions that did not follow the exact template, we had to manually align them. For UX101, past instructor and peer feedback were offered in correspondence with the rubric items. Solution segments and feedback offered to the solution were thus automatically matched. From this step, the solution file is reorganized and saved in a local database, an excerpt of which is shown in Table 3.2.

This manual checking step is a limitation of UpGrade’s current workflow. Potential ways to mitigate this when applying UpGrade in practice include: (i) logging assignment solutions using online forms where structures are predefined, eliminating the need of post-hoc segmentation and metadata association; (ii) abandoning falsely templated solutions when there is a large pool of existing solutions to source from; and (iii) applying advanced approaches to automatically align with the template to minimize the manual checking effort.

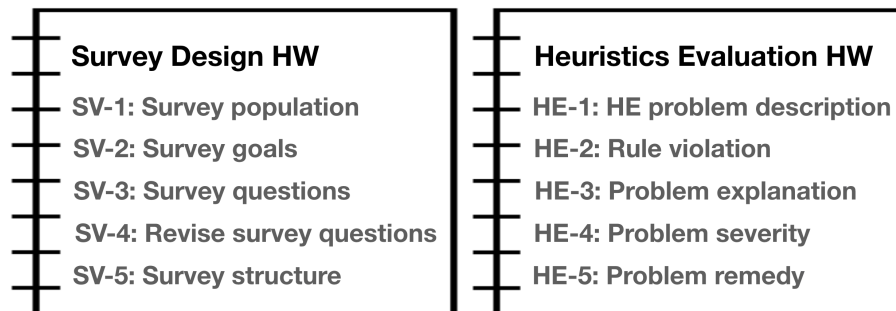


Figure 3.3: Rubric items for open-ended assignments on the topic of Survey Design and Heuristic Evaluation.

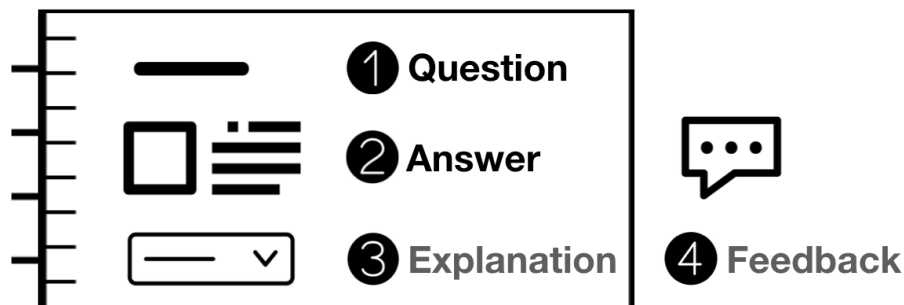


Figure 3.4: Four components used in UpGrade question schemata.

HW	Rubric item	Answer	Image	Feedback
Survey	Survey Population	The population of the survey are undergrads...	image_1	Quality of sleep is important to...
Survey	Survey Goals	How do people start to form their ideas of what...		'Do students share goals?'...
HE	Description	The user chooses to buy...	image_2	
HE	Violation	User control and freedom		
HE	Explanation	The only options the site...		
HE	Severity	Severity level 3 - major...		
HE	Remedy	A simple fix with no ...		

Table 3.2: A data excerpt produced by the segmentation step: past assignment solutions were segmented into sections based on the assignment rubric. Instructor and peer feedback was associated with solution segments when available.

3.4.3 Question Creation

We define four components *Question*, *Answer*, *Explanation*, and *Feedback* (Figure 3.4) to form question schemata in UpGrade. *Question* is a question asked in an open-ended assignment, *e.g.*, what are the goals of this survey. In doubly open-ended assignments, students may self-define a *Question*. *Answer* is a past student’s answer to a *Question*. Typical open-ended assignment solutions are composed of many *Question-Answer* pairs. In some assignments, students are required to offer *Explanation* to their answers. For assignments that have been graded, instructor or peer *Feedback* are also collected. Depending on the available data sources, instructors will (i) select a question schema and (ii) specify which sections should be placed into each component in the schema. Examples are given in Figure 3.5. With the segmented solutions produced (Table 3.2) and instructor-specified schemata, UpGrade then creates multiple-choice questions automatically. We introduce three question schemata we have defined and explored.

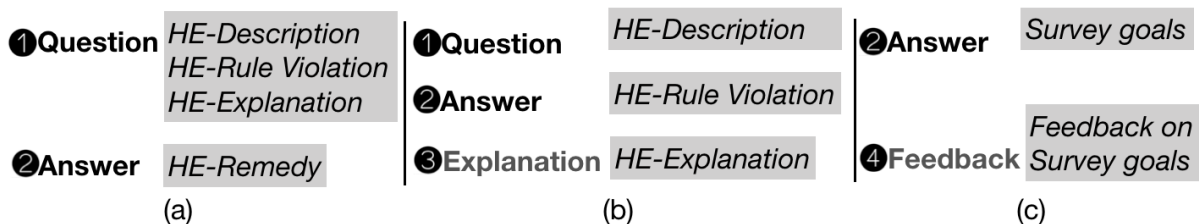


Figure 3.5: Example instantiations of the three UpGrade question schemata: (a) Question-Answer, (b) Question-Answer-Explanation, and (c) Answer-Feedback.

Question-Answer Schema

This schema defines a question with the components *Question* and *Answer*. In the example shown in Figure 3.5 (a), three solution segments including heuristic problem *Description*, *Rule Violation* and *Explanation* were used as the *Question*. *Remedy* of the problem was used as the *Answer*. The distractors were selected from the pool of *Remedy* that were written for other problems. The example question shown in Figure 3.6 displays a heuristic problem, and asks question takers to select a remedy that would fix the problem.

Question-Answer-Explanation Schema

When *Explanation* is available as a data source, it can be used to offer informative real-time feedback in the created question. This schema defines a question with the components *Question*, *Answer*, and *Explanation*. As shown in Figure 3.5 (b), the heuristic problem *Description* was used as the *Question*, and the *Rule Violation* was used as the *Answer*. The distractors were selected from the pool of *Rule Violation* for other problems. The example question shown in Figure 3.7 describes an interaction scenario of a website, and asks the question taker to identify which heuristic rule is violated. Since the original author offered an explanation on why the rule was violated, the corresponding *Explanation* is used as feedback to the question taker.

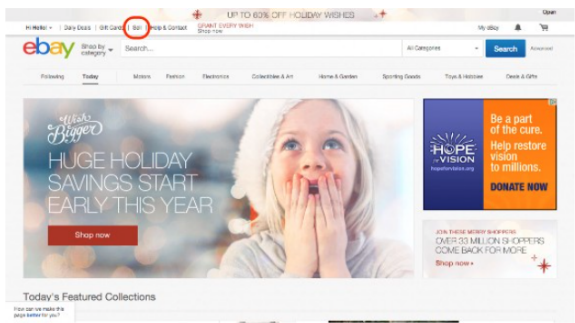
We present another example instantiation of this schema, when there are multiple iterations over a solution. In the survey design assignment, past students designed survey questions, revised them and explained why they made the revision. With this schema, draft 1 of a survey question was used as the *Question*, revised version of the survey question was used as the *Answer*. Figure 3.7 shows an example question created. Both versions of the survey question are displayed, and it asks question taker which version is better. Since the original author explained why they made the revision, the corresponding *Explanation* was used as feedback to the question taker.

Answer-Feedback Schema

This schema can be used when past instructor or peer *Feedback* is collected. Since *Feedback* points to prior students' misconceptions and common errors which may repeatedly happen with a new group of students, they can be a good source for creating questions. This schema defines a question with the components *Answer* and *Feedback*. As shown in Figure 3.5 (c), past students' solution of *Survey goals* was used as the *Answer*, and the feedback offered to this solution was used as *Feedback*. Distractors were selected from the pool of *Feedback* that have been offered to other solutions. The example question shown in Figure 3.9 displays past students' written solutions of survey goals and asks question takers to select which feedback would apply to each solution.

In this running example, after the segmentation step, we sat down with the UX101 instructor for about two hours in total to decide which question schemata to use and specify the sections to be used in each schema (the same process as shown in Figure 3.5). We asked the instructor to pick a schema for each rubric item to make sure UpGrade creates multiple-choice questions that cover the full scope of its open-ended assignment counterpart (Table 3.3). With the instructor-specified schemata, UpGrade creates multiple-choice questions automatically. For example, for HE-1, the specified schema is Q-A-E, also shown in Figure 3.5 (a). For every (*Description*, *Rule Violation*,

Below is a heuristics problem of Ebay identified by a previous student. Please examine the screenshot, read the problem identified and answer the question below.



Which of the following options do you think is the best fix to the problem shown on the left?

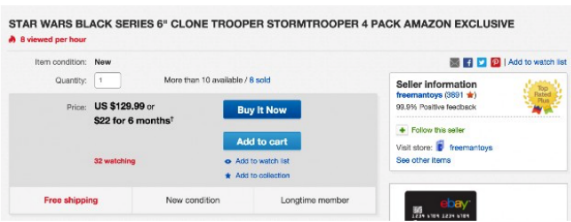
- The easiest way to fix this issue would be to have a section that highlights based on which box has an error. The best way would be to have additional text next to the highlighted section on how to fix the problem.
- Easiest way to fix would be changing the color of links in the bar to be blue instead of looking like black text. This would not disturb the color scheme on the page and doesn't seem to have any apparent tradeoffs. The best fix would be to place the link by a colored sell button and place it prominently maybe on the top-left or top-right of the page.
- A fix may be to delete irrelevant bars. They can do a user testing to see which navigation bar the users use more and get rid of the less popular one.

Walk-through of the problem:
 To reproduce: Open homepage. Try to find selling link. Get lost in content. For an online marketplace which relies on sellers, the selling feature is not prominent on the homepage. It took a couple of scans of the whole page to discover the selling link on the page and finally had to do a Cmd+F. It doesn't even look like a link.

Upgrade Feedback:
 The remedy you selected does not match this problem. The correct match is option #2 for this problem.

Figure 3.6: An example question created by UpGrade using the Question-Answer schema.

Here is a heuristics problem of Ebay identified by a previous student. Please examine the screenshot and answer the question based on your best knowledge.



Which one of the following Nielsen's heuristics rules does it violate?


- Flexibility and efficiency of use
- Consistency and standards
- Help users recognize, diagnose, and recover from errors
- User control and freedom

Upgrade Feedback:
 Sorry the better answer would have been: **User control and freedom**
Here's why according to a previous student. What do you think?
 When a user goes to an item's page and clicks on the 'Buy It Now' button, he/she is directed to a page that has two options: 'Commit to buy' and 'Continue with PayPal Credit.' There is no back button for users who may have accidentally selected 'Buy It Now' on the previous page. There should be an undo option so that, in situations such as this one, the user can leave the unwanted state without any trouble.

There is no back button after selecting "Buy It Now"

Figure 3.7: An example question created by UpGrade using the Question-Answer-Explanation schema.

Your classmates are asked to design a survey to better understand the service/user experience/mechanisms of UHS (University Health Services). They first came up with an initial list of questions and did pre-testing with potential respondents and came back to revise their survey. Please find below two versions of a survey question and select which one is better.

 **Upgrade Feedback:**
Great! You and a previous student both thought the **Version 1** is better.
Here's what (s)he said, do you agree?
For the question 'Are you addicted to alcohol?' it is changed to 'Do you have a history with alcoholism?' because people sometimes may not be aware that they are addicted.

Version 1
Do you have a history with alcoholism?
A. Yes B. No

Version 2
Are you addicted to alcohol?
A. Yes B. No

Which one is a better survey question?

Version 1
 Version 2

Figure 3.8: An example question created by UpGrade using the Question-Answer-Explanation schema (Revision variation).

Your classmates are asked to design a survey to better understand the user experience of UHS (University Health Services). They need to decide what research questions they want to get answered from the survey. Below are two student solutions and a list of peer feedback offered to the solutions.

Student A Solution: _____


Student B Solution: _____

Peer Feedback (match feedback with the above solutions):

1. Yes, I mean it makes sense to target all users of health services and understand this population because this is the population that is actually using HS. You are also trying to get the demographics of this population, comparing them to graduate students. I would say your goal for this survey is to get a general understanding of the demographic that uses HS and how they generally feel about their health and the university's health care services.
2. In survey purpose I'm a bit confused as you go from student life in freshman year(or possibly more) to talking about how UHS's services could be better explained through just teaching an RA. I feel this is more a Design Idea rather than a survey goal. The population however seems well justified from the rationale given.
3. One of your goals is to measure how often one gets sick, feels chronically stressed and sleeps during weeknights. However, your survey populations are determined by their exercise frequency. So maybe there is some disconnection between your goal and your population

Which feedback of the above 3 did Student A get?

1
 2
 3

 Great! You picked the same feedback given to this solution.

Which feedback of the above 3 did Student B get?

1
 2
 3


 The feedback you selected does not match this solution. The correct match is option #2.

Figure 3.9: An example question created by UpGrade using the Answer-Feedback schema.

Rubric	Schema	Description
SV-1	A-F	Match instructor feedback to student writing of survey population
SV-2	A-F	Match instructor feedback to student writing of survey goals
SV-3	A-F	Match instructor feedback (issues/suggestions) to each survey question
SV-4	Q-A-E	Compare original and revised question (with UpGrade feedback: student explanation on why they made the revision)
SV-5	A-F	Match instructor feedback to student design of survey structure
HE-1	Q-A-E	Decide which heuristic rule is violated in the problem (with UpGrade feedback: student explanation on why it violates the rule)
HE-2	Q-A	Match severity rating to a student-constructed heuristic problem
HE-3	Q-A	Match potential remedies to a student-constructed heuristic problem
HE-4	Q-A	Match potential tradeoffs to a student-constructed heuristic problem and its remedy

Table 3.3: Course instructor specified a question creation schema for each rubric item in the assignment.

Explanation) tuple, a question entry is created by selecting three distractors from the pool of *Rule Violation*. The questions produced by UpGrade are saved in a .csv file.

We built a prototype system with Django to render the questions. The front end of the prototype system looks similar to the interface as shown in Figure 3.6-3.9. With one year of past students' solution, UpGrade created large quantities of multiple-choice questions. The number of questions created for each rubric item is shown in the Space column of Table 3.4.

3.5 Classroom Experiment of UpGrade

We conducted a two-week experiment in a college-level HCI course to evaluate UpGrade in comparison with traditional open-ended assignments.

3.5.1 Crossover Experiment Design

We conducted this study in the Spring 2017 offering of the UX101 course, with 28 students enrolled. The course covered one topic (*i.e.*, research method) per week. Instructional activities on each topic included required readings, a 1.5-hour lecture, an open-ended assignment, and a 1.5-

Rubric	Trial	Pool	Space	Rubric	Trial	Pool	Space
SV-1	3	18	96	HE-1	30	70	478
SV-2	3	9	96	HE-2	10	70	478
SV-3	30	40	NA	HE-3	15	70	478
SV-4	10	11	NA	HE-4	5	27	91
SV-5	4	8	96				

Table 3.4: Space: number of questions created in total; Pool: number of questions used in the experiment; Trial: number of questions presented to students in each trial.

hour section. We divided students into two groups, Group A and Group B. Both groups of students did the same regular learning activities (readings, lectures, sections). The only difference was the type of assignment they did. For the topic of Survey Design, Group A worked on the traditional open-ended assignment, and Group B worked on UpGrade-created assignment. Similarly for the topic of Heuristic Evaluation, Group A worked on the UpGrade-created assignment, and Group B worked on the traditional open-ended assignment. Students were given about 7-10 days to finish each assignment.

For students working on UpGrade-created assignments, they logged in to a web-based system with their school ID and completed the assignment online. Student grades on this assignment were determined by how many questions they got right. In the system, students could navigate to different modules to work on the questions in that module. Modules align with the rubric items of the open-ended assignment (Figure 3.3). For each module, UpGrade produced a large question space. We ranked past student solutions by grade and selected high quality ones to be used in the experiment. The column of Pool in Table 3.4 indicates the number of questions used in the experiment on each module. Students had unlimited number of attempts at each module, allowing them to work repeatedly on the modules until they achieved a satisfying score. For each trial of a module, Trial number of questions were selected from the Pool (Table 3.4), giving students different learning opportunities in each trial.

3.5.2 Learning Outcome Measure

We administered a quiz on each topic in class as the learning outcome measure after each assignment was due. The quiz contained 8-12 questions, including both multiple-choice and open-ended questions. To counterbalance, each quiz item had two formats: an open-ended format, and a matched multiple-choice format. For example, a quiz item asked students to identify the design issue of a survey question. The multiple-choice form of the quiz item gave four options for students to choose from (*e.g.*, “leading question”, “asking about averages”), and the open-ended form gave a blank for students to fill in. In another example, the quiz item asked students to revise a survey question. The multiple-choice form of the quiz item gave four candidate questions for students to choose from, and the open-ended form asked students to revise the question in a text box. By varying the format for each quiz item, two variations of the quiz were created. Both variations had half multiple-choice and half open-ended questions. Students were randomly assigned to one of the variations.

The quiz was designed in collaboration with the course instructor to make sure it aligns with

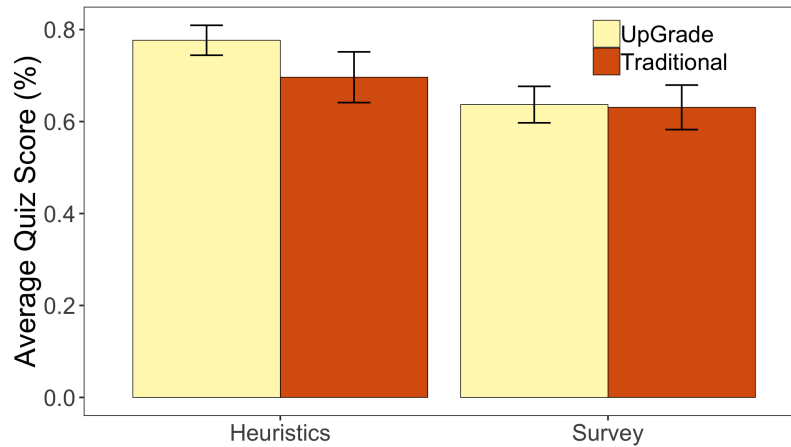


Figure 3.10: Student average quiz score in percentage by condition and content with standard error bars.

the course objectives. We also conducted a think-aloud session with a domain expert and made sure that the expected answers aligned with the experts' responses. There are 4 variations of each question item, so that there are only 7 data points for a given set of 8 question items. Thus it is hard for us to perform reliability tests on this dataset.

3.6 Experiment Results

Learning outcomes were analyzed in a *Condition* (UpGrade-created Multiple-choice vs. Traditional Open-ended) by *Content* (Heuristic Evaluation vs. Survey Design) repeated measures ANOVA. Results indicated a significant main effect of *Content* ($F(1, 26) = 5.76, p = 0.02$), with no main effect of *Condition* ($F(1, 26) = 1.02, p = 0.32$) and no interaction effect. This suggests that students who did UpGrade-created assignment achieved equal learning outcomes in comparison with students who completed traditional open-ended assignments. Surprisingly, we see a trend suggesting that students from UpGrade condition may actually have performed better on the quiz than the Traditional condition, as shown in Figure 3.10.

In the in-class quiz, students were also asked to self-report the time they spent working on the assignments. We performed another repeated measures ANOVA analyzing assignment completion time by *Condition* and *Content*. Results indicated a significant main effect of *Condition* ($F(1, 24) = 6.55, p = 0.017$), with no main effect for *Content* ($F(1,24) = 0.001, p = 0.97$), and no interaction effect. The average assignment completion time by *Condition* and *Content* is displayed in Figure 3.11.

Overall, when students did the UpGrade-created assignment composed of multiple-choice questions instead of the traditional open-ended assignment, there was a 28% reduction in assignment completion time, from an average of 6.34 hours ($SD = 3.03$) to 4.56 hours ($SD = 2.63$). The significant results show that this time reduction is substantial. Despite spending less time, students achieved equal learning outcomes. Moreover, the trend in learning outcome even favor the UpGrade condition (Figure 3.10). Further, UpGrade removed the need of manual grading

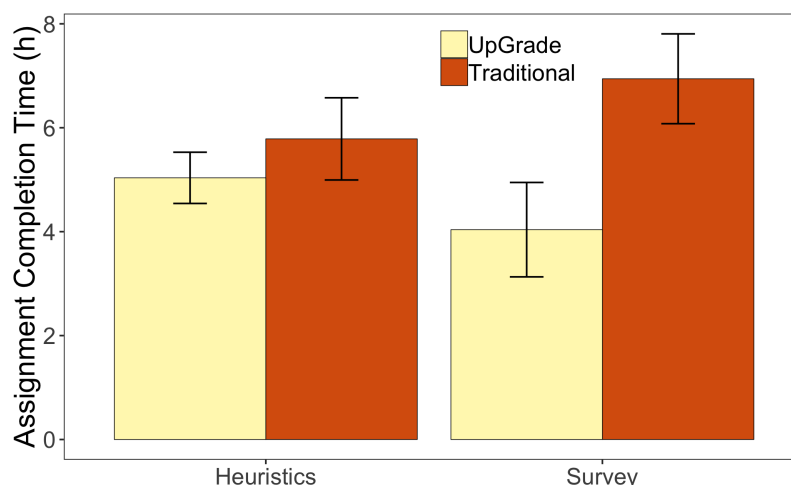


Figure 3.11: Student average assignment completion time in hours by condition and content with standard error bars.

effort from instructors and TAs.

3.6.1 User Experience and Feedback

To better understand user experience and get user feedback to improve UpGrade, we conducted a subsequent interview with the instructor and an in-class interview with the participating students. The instructor liked this approach in that students’ grades were all computed automatically, saving substantial efforts of grading and offering feedback. The instructor further expressed concerns that many students did not do well in the open-ended assignment. *“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 the assignment.”* Additionally, the instructor envisioned future practice where students got to practice with UpGrade first to learn the skills before they went off to generate new content.

Students gave feedback freely during an in-class group interview at the end of a lecture session. Participating students brought up usability issues of UpGrade and suggested ideas for improving the questions in the future. One student commented on the UpGrade heuristic evaluation assignment: *“It’s hard to understand the interaction scenario captured by the previous student from a static screenshot. Sometimes we have to guess the intention of the original author.”* Another student suggested *“In the automated feedback, it gives a more detailed description of the scenario. It’ll be helpful to move some of those texts up to the question stem to illustrate the screenshot.”*

The classroom experiment demonstrated UpGrade’s success in saving instructors’ grading time and reducing students’ time to complete a required assignment without sacrificing learning. Subsequent interviews with instructor and students suggested ways to enhance question quality. Though concerns that are inherent to the learnersourcing input (*e.g.*, image quality, text formats) requires more substantial effort to improve, which we will discuss in future work, there is a huge potential to select high quality items taking advantage of the large question pool.

3.7 Discussion, Limitations and Future Work

In this section, we discuss the limitations and potential future directions to enhance the question quality and the learning benefit of UpGrade.

3.7.1 Structured Text Data Logging

UpGrade enables the creation of multiple-choice questions from existing data, saving instructors' efforts to manually construct materials. One important step in UpGrade's workflow (Figure 3.2) is to segment existing open-ended solutions into sections based on the assignment rubric. In our experiment, manual effort was required in segmenting the existing solutions. A better approach would have been logging assignment data hierarchically through digital forms. This would eliminate the need for UpGrade to segment assignment texts.

3.7.2 UpGrade As A Primer To Open-ended Assignment

One potential risk of UpGrade is that it does not allow students to produce content as they would normally do in open-ended assignment. On the one hand, students would not engage in successful content creation before they have mastered the required competence. On the other hand, we do not argue UpGrade should replace traditional open-ended assignment. In cases where the goal is to develop mastery towards certain knowledge and skills, UpGrade can be used alone; in other cases where the goal involves content creation, *e.g.*, projects to be included in portfolios, UpGrade can be used as a primer to open-ended work to prepare and scaffold students towards higher quality content generation.

3.7.3 Quality Control and Quality Enhancement

We propose three directions for better quality control in UpGrade. *(i)* Employ active learner-sourcing. The current workflow of UpGrade completely relies on existing learner-generated written content, without intervening the content production process. Future work might explore interventions on the content production process [65] to support more active learnersourcing, *e.g.*, prompting students to document their thought processes while writing open-ended solutions may produce additional input for question creation in UpGrade. *(ii)* Employ NLP techniques to improve text clarify and select better distractors. For example, removing irrelevant texts from student solutions; add intelligence into the system in selecting distractors (similar distractors, abstract distractors, etc.) *(iii)* Develop an instructor review phase in UpGrade for instructors to review, revise, and select questions. Intelligent support can be provided to instructors while they are reviewing the questions, *e.g.*, highlighting the texts that may require clarification. This aims at better leveraging the capabilities of human and machine for high quality content production.

3.8 Conclusion

In this work, we contribute a novel learnersourcing approach, UpGrade, that creates multiple-choice practice questions with immediate feedback using prior student solutions to open-ended problems. An evaluation experiment demonstrated that students achieved indistinguishable learning outcomes in ~30% less time from UpGrade compared to traditional open-ended assignments, while at the same time eliminating the need for manual grading. UpGrade also incorporates a quality control method that prunes out low quality questions based on student performance data. With continued development, we envision a broader impact of UpGrade to generate high quality learning opportunities that easily scale up and benefit learners and education providers.

Chapter 4

Evaluation of UpGrade Using Open-ended Task Performance Measures

Following the first experiment (Chapter 3) in which we used quizzes that comprise both multiple-choice and open-ended questions as outcome measures, we were interested in seeing whether exercising through UpGrade (evaluation-type activities) could help students improve their open-ended work quality. This also helps us to re-imagine potential new take-home assignment models with UpGrade.

We ran this experiment in the fall of 2019, in a user research methods course offered at CMU. The course covers one research method each module. Our study was conducted on the module of analyzing usability findings from think-aloud studies and the module of designing storyboards for speed dating studies.

There are three major research goals for this study. First, we would like to investigate whether UpGrade-created multiple-choice questions can help students exercise the skills relevant to generating responses in authentic and complex open-ended tasks. Second, the study design corresponds to a new take-home assignment model where students exercise with evaluation-type activities first and then go off to generate open-ended content. On the one hand, we are concerned about the quality of the product produced. On the other hand, we are concerned about the total time it took for students to complete the assignment. Our hypothesis is that if students get some discrete practice first, it could actually save their time in doing the subsequent open-ended tasks. Third, we implemented the UpGrade technique following the steps as outlined in Figure 4.9, with co-design sessions with instructors in step 3 and 5. The goal is to understand what are instructors' preferences and challenges when coming up with question schema and revising questions, to better inform the design of an end-user interface for UpGrade.

4.1 Study Design

In each module, there is a 1.5-hour lecture, and the assignment is announced at the end of the lecture, the assignment is due in a week. There is a TA-led section during this week for students to ask questions and discuss their work with the TA. The class has five sections. Three sections were assigned to Group A, and two sections were assigned to Group B. The two

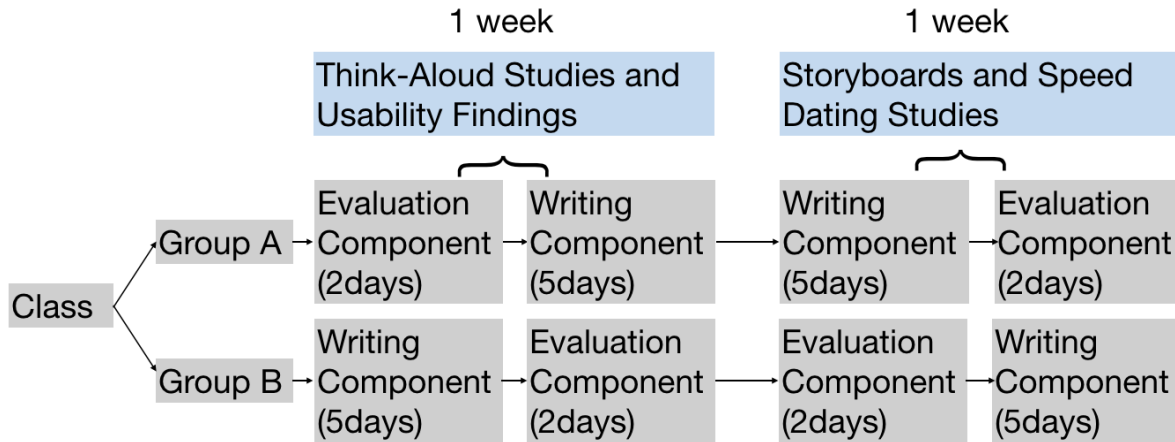


Figure 4.1: Study 2 Design:

experimental assignments contain two components, an evaluation component, in which students complete UpGrade-produced multiple-choice questions through a web-based system, and a writing component, in which students complete the projects in groups. Other parts of the class proceed as usual. In the first week, students are learning about think-aloud protocols and summarizing usability findings from them. The open-ended assignment is a week-long group project. In the UpGrade condition, we give students an exercise composed of 20 multiple-choice questions produced by Upgrade before the students work on their projects. In the second week, the two groups switched when learning about storyboard and speed dating techniques. We use the quality of students' open-ended projects as the learning outcome measure.

4.2 Study Materials

We show the study materials for the module of storyboards.

Assignment Announcement on Canvas The assignment announcement is as shown in Figure 4.2. Students were also asked to complete a survey at the end of the assignment to report time spent doing different components of the assignment. Detailed prompts for the evaluation and writing component are shown in Figure 4.3 and Figure 4.4.

Example Questions In the UpGrade condition, students log in to the system to complete the activity. There are 4 sections here corresponding to the 4 question schemata the instructor has created. Section 1 is about the alignment between storyboard and user need. An example question is shown in Figure 4.5. Section 2 is about writing lead questions for storyboards. An example question is shown in Figure 4.6. Section 3 is about evaluating the riskiness of storyboards. An example question is shown in Figure 4.7. Section 4 is about evaluating the quality of storyboards,

Assignment

- We are doing an experimental assignment on this topic that aims at improving students' learning experience in the course. The assignment 2.4 has three components
 - The Evaluation Component contains multiple-choice questions in which you will evaluate past examples of Storyboards.
 - The Writing Component is an open-ended assignment in which each of you will design storyboards and conduct speed dating with it.
 - The Peer Review Component asks you to provide feedback to each other and comprehend the feedback you receive.
- We will use the same grouping as we did for the previous assignment. Group 1 will do Writing Component first. Group 2 will do Evaluation Component first. Both groups need to submit your peer reviews by Sunday. You will have two deadlines for this assignment. You will see your corresponding deadlines in the Assignment turn-in link. **Please follow the required sequence here to meet both deadlines.**
- Group 1 (Section A, Section D, Section E):
 - [Project Deliverable 2.4: Writing Component](#) due Friday (Nov. 15)
 - [Project Deliverable 2.4: Evaluation Component](#) due Sunday (Nov. 17)
 - [Project Deliverable 2.4: Peer Review of Speed Dating Session](#) due Sunday (Nov. 17)
- Group 2 (Section B, Section C):
 - [Project Deliverable 2.4: Evaluation Component](#) due Tuesday (Nov. 12)
 - [Project Deliverable 2.4: Writing Component](#) due Sunday (Nov. 17)
 - [Project Deliverable 2.4: Peer Review of Speed Dating Session](#) due Sunday (Nov. 17)

Time Completion Survey

Upon completion of the assignment, please answer this survey on how much time the assignment takes. This will only take 1 minute. This is to help improve the design of the course!

<https://forms.gle/F6dKivLpgs1pNGYm9>

Figure 4.2: Study 2 Materials: Assignment Announcement for “Ideation, Storyboarding and Speed Dating ”

1. (Individual) Complete questions from website:

Go to website: luoyang.lti.cs.cmu.edu:3000

and complete the questions there. The final score will be determined by how many questions you get correct.

You have unlimited attempts until the deadline [see below] to complete this. We will keep the highest score as your grade. After the deadline, you can still review the materials from this website.

Figure 4.3: Study 2 Materials: “Ideation, Storyboarding and Speed Dating” Evaluation Component Prompt

3. Create Storyboards

1. Analyze the breakdowns/needs, approaches, ideas and unanswered questions that were raised in the Crazy 8s game to identify the greatest areas of uncertainty and risk. As a team, generate a list of user needs from this analysis.
2. Looking at the list of user needs, each team member should select one user need for which they will create a set of storyboards. Each team member will then sketch three distinct scenarios based on the one user need:
 - one scenario is the "safe" way to meet the user's need
 - the second scenario is progressively riskier
 - the third scenario should be intentionally "out there" and intended to make the user uncomfortable and test their boundaries.
3. Each storyboard should have 3 panels:
 - First panel sets up the scenario, e.g. how each need arises in daily life
 - Second panel illustrates the idea and how it solves the need
 - Third panel demonstrates the outcome (the improved experience)
4. As part of your document, generate a cover page for each storyboard that lists: the user need, the lead question, and several bullet points for discussion with users. This will NOT be shown to users.
5. Storyboards **SHOULD**:
 - Focus on situations where it's easy for participants to imagine themselves.
 - Show people in specific contexts interacting with the intervention.
6. Storyboards **SHOULD NOT**:
 - Draw attention to specific technical solutions that distract users from the focus on the need and unintentionally dominate conversations.

Figure 4.4: Study 2 Materials: “Ideation, Storyboarding and Speed Dating” Writing Component Prompt

through matching storyboards with previous instructor feedback. An example question is shown in Figure 4.8.

4.2.1 Implementation of UpGrade

We followed the 5-step workflow as shown in Figure 4.9. In step 1, we collected students' submitted assignments data in the past. In step 2, we segmented the solutions using keyword matching; an example is shown in Figure 4.10. We did a sanity check, in the end, to make sure the segmentation is done meaningfully. In step 3, we did this through a one-hour meeting with the instructor. During the meeting, I bring the materials, including the original student submission files and the segmented data and discuss with the instructor about possible question schemata. For example, the instructor would talk about the learning objectives for the module and suggest question schemata. In step 4, using the specified question schemata, I run the algorithms to reorganize the segmented responses and select distractors to create multiple-choice questions. In step 5, we also did this through a one-hour meeting with the instructor. During the meeting, we go over the question pool and the instructor would make edits when necessary.

4.2.2 Analysis Methods

There are two outcome measures we used here. First, the quality of the student's open-ended work. Second, the completion time for the assignment. To gauge the quality of the student's

Section 1

Design Storyboards Based
on User Needs

Section 2

Write Lead Questions

Section 3

Safe and Risky Storyboards

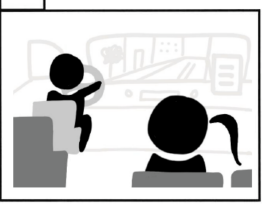


Section 4

Is This a Good Storyboard?

In this activity, imagine your peers are designing Storyboards to improve users' transportation experience around CMU. They first identified a list of user needs and then designed Storyboards to address the needs. Below, you will see Storyboards created by them. Your task now is to assess the **user need** each Storyboard addresses.

Question 1 out of 6

Below is a Storyboard to improve users' transportation experience around the CMU campus:

SHOT #	SHOT #	SHOT #
		
Jen is sitting in the front row of the shuttle, diagonal from the driver.	Jen sees the display that signals that the driver is currently "in the mood for conversation."	Seeing the display, Jen says hi and before she realizes, they begin sharing stories about their lives. Jen creates a new, unexpected connection during her shuttle ride.

Which of the following best describes the **user need** this Storyboard is designed for?

- Riders need a reliable centralized channel with all shuttle information.
- Students and drivers need a more personal relationship with each other.
- To foster better communication and reduce barriers with drivers.
- Riders need to know the current capacity of the buses to make better decisions.

Optional: leave additional comments

Check Answer

Figure 4.5: Example question of Section 1: asking students to match a user need for a shown storyboard

Section 1

Design Storyboards Based
on User Needs

Section 2

Write Lead Questions

Section 3

Safe and Risky Storyboards

Section 4

Is This a Good Storyboard?

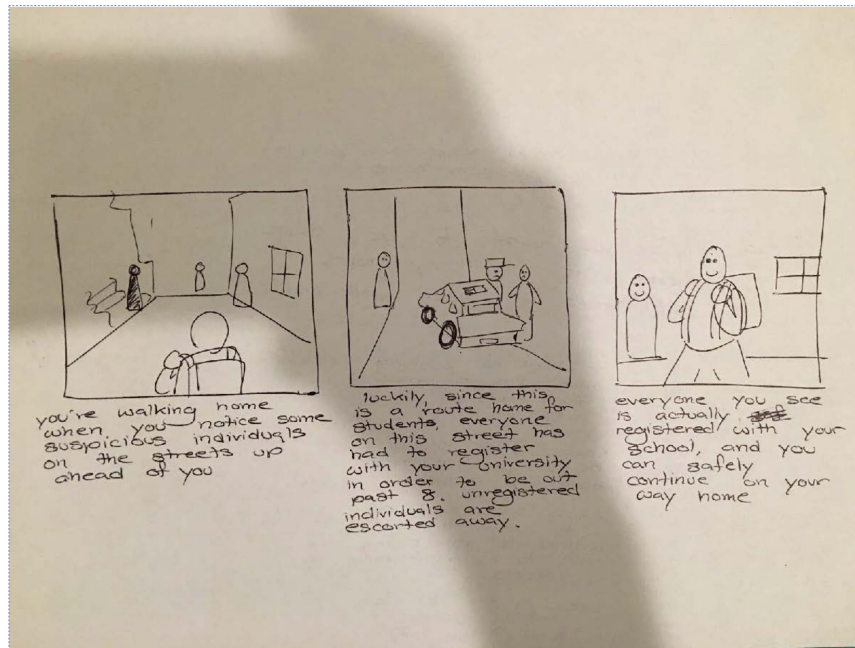
In this activity, imagine your peers are designing Storyboards to improve users' transportation experience around CMU. They first identified a list of user needs and then designed storyboards to address the needs. In the next step, they will use these Storyboards to conduct Speed Dating session with users. They wrote Lead Questions for the speed dating sessions. Your task now is to evaluate which **Lead Question** is the best for each Storyboard.

Question 1 out of 5

This Storyboard addresses the following user need:

Avoid suspicious people on your way home

Storyboard:



Which of the following would be a good Lead Question for Speed Dating using this Storyboard?

- Do you want to meet other shuttle riders and are unsure of who also wants to talk?
- Do you need exact navigation to classrooms on campus?
- Have you ever realized last minute that you're running late? how did you take steps to make sure you made your flight on time?
- Have you ever worried about suspicious individuals on your walk home?

Optional: leave additional comments

Check Answer

Figure 4.6: Example question of Section 2: asking students to match a lead question for a shown storyboard

Section 1

Design Storyboards Based on User Needs

Section 2

Write Lead Questions

Section 3

Safe and Risky Storyboards

Section 4

Is This a Good Storyboard?

In this activity, imagine your peers are designing Storyboards to improve users' transportation experience around CMU. They first identified a list of user needs and then designed Storyboards to address the needs. For each user need, they created three storyboards that are "safe", "riskier", and "out there" ideas. Your task is to evaluate whether each set of the three storyboards follows the **progression of increasing riskiness**.

Question 1 out of 5

This set of Storyboards addresses the following user need:
More personalized information on shuttle service

Safe design:

Shuttle Website Search Feature

1. SEARCH DESTINATION
Q. ADDRESS, MENU, etc.

SUGGESTED
ROUTE A
POINTS & MARKERS STOP

2. SHUTTLE ONBOARDING

Search CMU Shuttle Website for shuttle info Click on First Timer button Quick Search feature to find out the optimal route & stop

Riskier design:

Shuttle Stop Interactive Display

RESULTS

ROUTE A (HARVARD CENTER MARKET) [24min] LEAVING IN 5:15 (5min)

ROUTE AB (34min)

ROUTE B (44min)

Approach the interactive display at shuttle stop Use the display to search for stops Display search results in a list

"Out there" design:

Auto Suggestion with SIO Integration

WE FOUND A NEW SHUTTLE ROUTE FOR FOOD!

CMU SHUTTLE SERVICE COMMUNICATED TO BUS

Hi _____

YOU CAN NOW TAKE ROUTE AB TO HOME FOR 25 MIN!

Update SIO address as required Receive Notification from school email Message indicate the optimal route and stop for the new address

Does this set of three storyboards follow the progression of safe, riskier, and "out there" ideas?

- Yes
- No

Sorry, the three storyboards do not show a progression of riskiness in the design. This shows a progression of more complicated technologies. But it doesn't push the boundaries on uncomfortable human behaviors.

Optional: leave additional comments

Next Question

Figure 4.7: Example question of Section 3: asking students to evaluate whether the set of three storyboards follow a progression of riskiness. The feedback here is written by the instructor post-hoc, in step 5 of the workflow (shown in Figure 4.9

Section 1Design Storyboards Based
on User Needs**Section 2**

Write Lead Questions

Section 3

Safe and Risky Storyboards

Section 4

Is This a Good Storyboard?

In this activity, imagine your peers are designing Storyboards to improve users' transportation experience around CMU. They first identified a list of user needs and then designed Storyboards to address the needs. In the next step, they will use these Storyboards to conduct Speed Dating session with users. They wrote Lead Questions for the Speed Dating sessions. You will read examples of Storyboards and associated Lead Questions. Your task now is to evaluate the quality of the examples and **select a piece of feedback** for each.

Question 4 out of 4

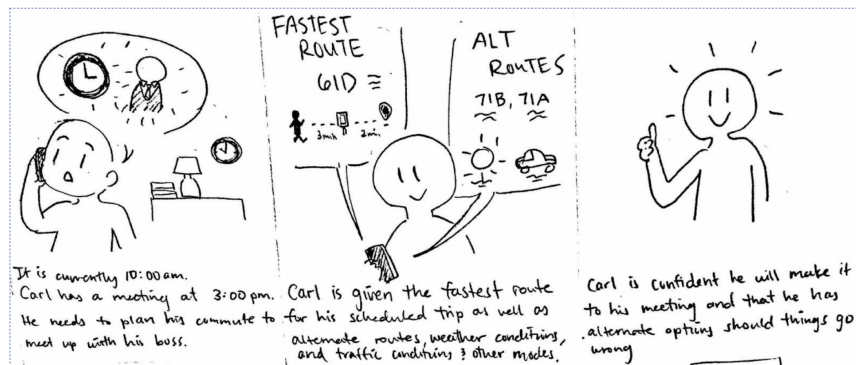
This storyboard addresses the following user need:

Riders want to plan a trip for a specific time and purpose in the future.

Lead Question:

What kinds of information (time, weather conditions, traffic, etc.) do you consider or find useful while planning a trip for later in the day/the future?

Storyboard:



Which of the following is the most informative feedback you could offer to this Storyboard?

- This is a good example of Storyboard.
- Storyboards shouldn't focus on specific screen designs.
- The lead question should evoke the participant to discuss a similar situation they may have been in.
- Storyboards should be designed based on users' needs rather than other stakeholders' needs.

Optional: leave additional comments

Check Answer

Figure 4.8: Example question of Section 4: asking students to evaluate the quality of a shown storyboard, through matching storyboards with previous instructor feedback.

Workflow of UpGrade

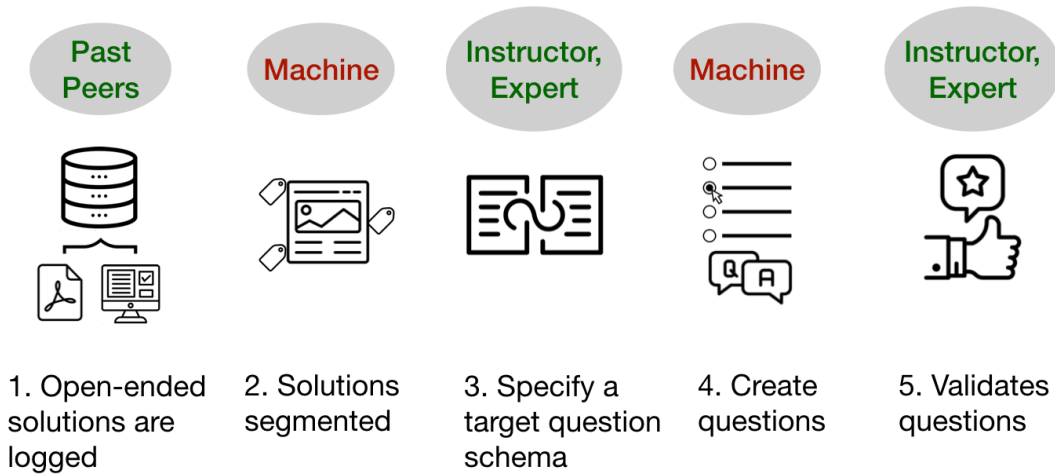
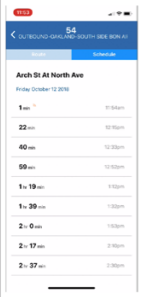


Figure 4.9: Five Step Workflow of UpGrade

No. TAI-1	Problem
Name: User not sure how to get stop arrival information for a bus in the future (other than the next bus coming).	
Evidence: When figuring out how to get to the Mattress Factory, user tried to click into the bus schedule to understand when the bus after the next would arrive near campus. He tried tapping multiple times and spent 45 seconds tapping on the arrival time for the bus after next. Note that the schedule and arrival times shown were for the stop near the Mattress Factory, not near campus.	
TA1MOV	
	
Explanation: User expected to get additional details about the overall bus schedule beyond the next bus – like what he was seeing on the “route” page, but for a future bus. He wanted this because the bus was currently near the Mattress Factory and showing 1 minute away, but he was trying to figure out when he should leave to catch the bus to the Mattress Factory. It is unclear whether he understood that the schedule reflected the bus arrival times at a specific stop.	
Severity or Benefit: Rating: 3 (Major)	
Justification (Frequency, Impact, Persistence, Weights): Frequency: This is a frequent issue as it impacts any user who wants to understand information about future bus arrivals past the next bus coming, and the app does not default information about that stop. Impact: This is difficult to overcome for a user because it requires knowing the stop that is closest to you, searching for that stop in the universal search, selecting the route that you intend to take, and navigating to the schedule. Persistence: Once a user figures out this sequence of steps, he/she can overcome it but it will likely take a significant amount of time using Taramsu before he/she is able to fully understand and overcome the confusion. How I weighted the factors: This is a frequent issue with a work-around, but the work-around requires multiple steps and not easy to discover on your own. For those reasons, frequency and impact outweigh persistence. Possible solution and/or trade-offs: Allow the user to click through each arrival time on the schedule to see a view similar to “route” for the corresponding bus time. This requires significant engineering effort to build another screen and UX effort to confirm that the information architecture still makes sense.	
Relationships: N/A	

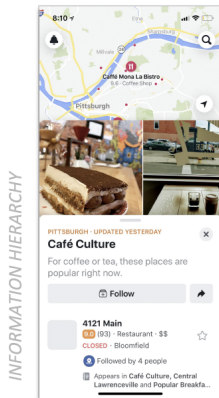


	A	B
1 ID	TA1-1	
2 Problem	Problem	
3 Name	Name: User not si	
4 Evidence	Evidence: When fi	
5 image	<pic:pic><pic:nvPic	
6 Explanation	Explanation: User	
7 Severity	Severity or Benefi	
8 Rating	Rating: 3 (Major)	
9 Frequency:	Frequency: This is	
10 Impact:	Impact: This is dif	
11 Persistence:	Persistence: Once	
12 weighted	How I weighted th	
13 solution	Possible solution :	
14 Relationships:		

Figure 4.10: Example of segmentation: through a keyword matching algorithm, the solution on the left is transformed into segmented responses on the right.

#8 Lack of information on location on event page

Individual Event Page



During 0:57 - 1:02, the user was confused on what to do next after selecting an event that interested her

Findings

- The user was confused and unsure about the event because she didn't know where the event was at and for what purpose
- From the recommended list, she selected "Café Culture" as it sounded interesting, but immediately left after 5 seconds as she was unsure of where the event was
 - Frequency:** Users are likely to experience this issue as a lot of the event pages do not have a clear information hierarchy for each of the hosted events such as addresses, contact, etc.
 - Impact:** It may be difficult for users to overcome this issue because the only actions that a user can take from here is either to follow the event or share the event to other friends, which isn't solving their needs on figuring out what the address of the event is
 - Persistence:** Users may be repeatedly bothered by it as it doesn't display the address of the event, instead it uses Facebook's integrated map which doesn't clearly show the users on what next steps to take to get to that event. Another user in [Slide 7] also struggled with this issue.

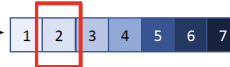
Recommendations

- Solution:** Include the address of the publicly hosted event and the point of contact for that event in case of help
- Trade-offs:** Linking another Facebook user's profile as a point of contact may cause privacy concerns

Based on UAR by Brad A. Myers & Bonnie John

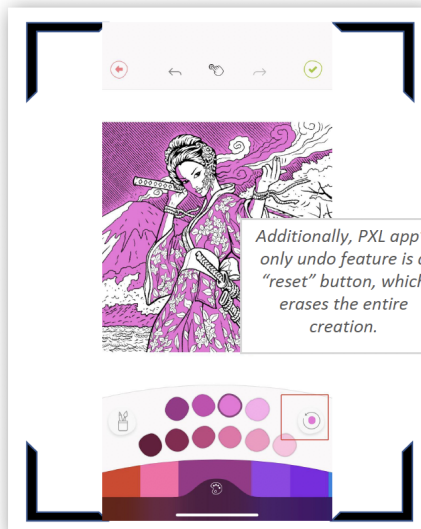
Figure 4.11: UFT Student Work Example 1

Severity Rating →



#6 Undo icon's function is not obvious to user

Colorfy drawing toolbar once user start drawing



Findings

- Evidence:** Users #4 & #5 couldn't find the undo button to erase their misplaced color when they wanted to eliminate the color of an area.
- Frequency:** Both users had problems figuring out how to fix their mistakes because the erase icon on the right is not clearly read as an undo button.
- Impact:** It is easy to overcome because once users know what the button does, they will be able to use it.
- Persistence:** It is a one-time problem that users will learn to overcome. It will only appear the first time the user tries to undo one of their colors.

Recommendations

- Solution:** Change the undo icon to a more commonly used icon, such as a left arrow or an eraser.
- Trade-off:** Users who haven't used similar software, like photo editing apps, may not know what an eraser icon does.

Based on UAR by Brad A. Myers & Bonnie John

Figure 4.12: UFT Student Work Example 2

#1: Convenient Enjoyment

User Need

People want easy access to enjoyment, and art when it is convenient

Lead Question

Do you ever wish you were more aware of the art in your community?

Discussion Points

- Do you tend to notice public art when you aren't looking for it?
- Would you like to have your attention drawn more to public art?
- Would you have any concerns about a location-based notification?

Figure 4.13: Storyboard Student Work Example 1 (Descriptions of the storyboard and relevant meta-data for speed dating.



Figure 4.14: Storyboard Student Work Example 1

Presenting Art Information by using AR

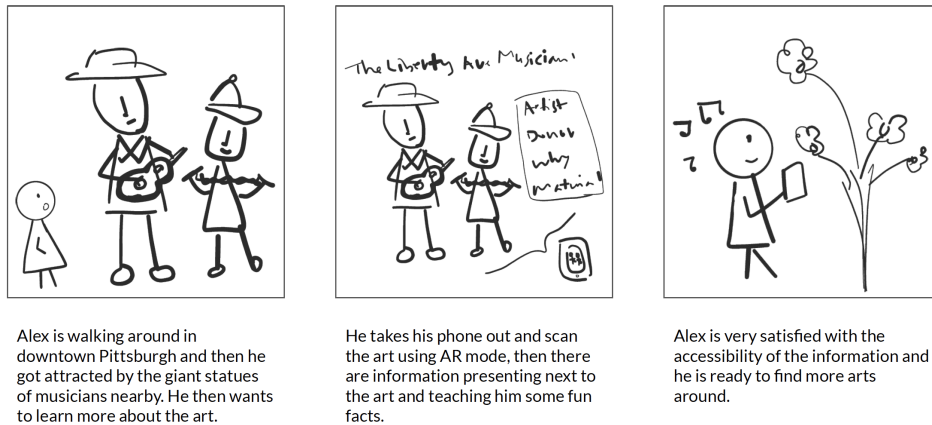


Figure 4.15: Storyboard Student Work Example 2

work, two instructors designed a very detailed grading rubric and checked inter-rater reliability before grading the entire assignment pool. Student open-ended work was randomized in order and named with random IDs before the instructors graded them, to make sure that the grading is blind to condition, independent and not affected by the sequence. After two instructors checked inter-rater reliability, one instructor graded all the Storyboard assignments (89 pieces), and the other instructor graded all the UFT assignments (187 pieces). Because each student was asked to write 2 UFTs and some students wrote more than 2, there were over 89×2 pieces in total. We computed the percentage score for each piece.

Figure 4.11 and Figure 4.12 show examples of student work on the *Usability Findings assignment* (summarize usability findings from think aloud studies). Figure 4.15, Figure 4.14 and Figure 4.13 show examples of student work on the *Storyboards assignments* (creating storyboards to prepare for speed dating studies).

4.3 Experiment Results

Storyboards quality significantly improved for students who used UpGrade beforehand

For the *Storyboard assignments*, among the 89 pieces submitted, 36 were in the UpGrade condition (who did the UpGrade-produced multiple-choice questions before doing the project), and 53 were in the control condition. With a Welch Two Sample t-test of *Percentage score on Condition*, we find that the UpGrade condition has significant higher grades compared to the control condition ($t=2.12$, $p=0.037$). The average grade for the UpGrade condition is 24.07 (83.0%), and the average grade for the control condition is 22.25 (76.7%). For the *Usability Findings assignments*, among the 189 pieces submitted, 117 were in the UpGrade condition and 72 were in the control condition. With a Welch Two Sample t-test of *Percentage score on Condition*, we see no difference between the two conditions on the total grade ($t=8.35$, $p=0.40$). The average grade for the UpGrade

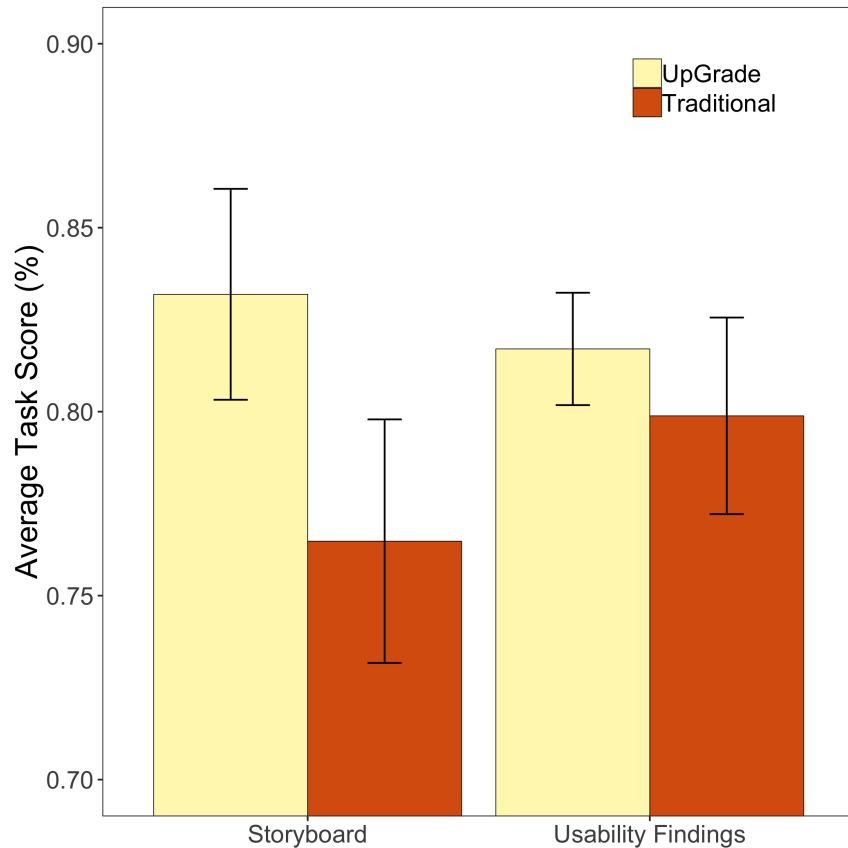


Figure 4.16: Average Percentage Score for both conditions on the two topics.

condition is 8.95 (81.4%), and the average grade for the control condition is 8.74 (79.5%). The average score for both conditions is displayed in Figure 4.16.

One possible explanation we see improvements on the storyboard assignments but do not see improvements on the usability findings assignments could be that, for the usability findings assignments, students first conduct think-aloud studies with real users and then summarize their findings from the respective think-aloud sessions. One thing we have observed while grading student assignments is that the quality of their usability findings reports could depend on the quality of the think-aloud studies that they performed beforehand. This may introduce more unobserved variance into the quality of students' open-ended work. Follow-up analysis such as targeted error reduction (e.g., Are all errors being made in the two open-ended assignments targeted in the MCQ?) could help us find more.

One limitation of the study design is that since all students put their individual work into the group assignment they turned in, we could not align individual student's work between the two topics. The crossover can only happen at a team level. There are 20 teams, with 10 teams in each condition. We did a repeated measures regression analysis at the team level, using the Percentage score as the dependent variables, the condition, topic and the interaction between the two as independent variables, and added a random intercept for each team. The result of this regression analysis is shown in Table 4.1.

	Model 1: Team Level Repeated Measures		Model 2: Individual Work	
	Coefficient Estimate	p-value	Coefficient Estimate	p-value
Condition (UpGrade)	0.07	0.13	0.063	0.056 .
Topic (Storyboard)	0.015	0.72	0.016	0.57
Condition (UpGrade)*	-0.05	0.51	-0.043	0.277
Topic (Storyboard)				

Table 4.1: Parameter estimates and p-value for the repeated measures regression analysis at the team level and at the individual level

Chapter 5

Iterative Design of QuizMaker

We followed an iterative process to design the authoring interface in UpGrade: QuizMaker. This includes an initial need-finding formative interview study to understand the challenges in designing instructional and assessment activities by instructors and learning engineers and an iterative design study working with instructors to apply UpGrade in their classes, examine what worked and what did not work, and summarize user preferences to improve the design.

5.1 Initial formative interviews

5.1.1 Methods

Four faculty members and three learning engineers from an R1 institution participated in this formative interview study. In the interview, I asked participants to describe their experiences in designing assignments and other course activities. The interviews were transcribed through an online transcription service. I then conducted a thematic analysis [18] and summarized the major themes emerging from our interview data.

In the next section, I summarize the major themes emerging from our formative interview study. Some themes were mentioned by both instructional designers (learning engineers) and instructors, and some were only mentioned by one stakeholder. I want to note here that the interview responses from the learning engineers were very consistent, and the interview responses from the instructors were more diverse. The goal of this initial formative interview study is to inform our system design and the themes should not be used as empirical findings.

5.1.2 Themes emerging from the formative interview study

Figuring out learning objectives is critical and takes time

In instructional design, figuring out the learning objectives of the class is critical and takes time. Instructional designers also need to break down learning objectives into specific skills, which is often an iterative process. Sometimes there are existing course materials, and what instructors do is to tweak based on the existing materials. Learning engineers also mentioned the challenge of not being domain experts in coming up with learning objectives, and sometimes instructors are

not available, and “having something to get started and have the experts refine them (P5)” could be an effective way.

Grading open-ended assignments is expensive

Our participants mentioned that grading open-ended assignments is “expensive”, and in one large course, “*the TAs only grade a subset of student submissions for each assignment (P7)*”.

Instructors and TAs may have different knowledge base

All instructors mentioned the challenges of having TAs grade and expressed concerns such as “They just don’t have the same knowledge base that I do. (P6)”

“Because you have to know more to see if this is right or wrong. Yeah. So the TAs are always assigned to grade, the ones that have a very small, like more application questions. It’s a sort of a small leak. The questions that are a bigger lead, I typically have to grade. (P6)”

Writing formative assessment is hard and time-consuming

In designing multiple-choice questions as formative assessments, most of our participants find writing them to be hard and time-consuming. “*So, writing the question is kind of the most hard, hardest part because there isn’t a particular way to go about it. (P3)*”

“*And mainly, I think because to write good ones, it’s very hard to make them. I sort of have a bias that it’s also to lower levels of Bloom’s taxonomy. But there are people who use very good questions that get to higher levels. They’re just extremely hard to write. It’s easy to write bad multiple-choice question. Yes. And so I don’t think it’s impossible, but I have not mastered that. (P6)*”

Coming up with examples and scenarios is hard

In particular, our participants find coming up with examples and distractors to be hard. “I might sit there for 15 minutes trying just to think of a scenario or an example. (P3)”

The learning engineers also mentioned it is important to collaborate with domain experts to identify plausible distractors. “. . . have to work with the subject matter expert. Like sometimes, if they’re writing questions, usually they’ll provide that and we might refine and ask why did they choose that distractor, i.e., where’s that coming from? But we have to, like learn enough of the content to do that and then work with them to say like, Oh, I put this because I think it’s this thing, can you check whether that’s a good distractor? (P2)”

Creating isomorphic tests, predicting the difficulty of questions is hard

Our participants found predicting the difficulty of the questions to be hard. “*If we were generating questions for, like a pre or post test, which was we were doing as part of that project, and making sure that it is same difficulty level, but without being the same question. It’s hard. It’s really hard. (P2)*”

Learning engineers also mentioned that they are not subject matter experts, making it harder for them to predict the difficulty. *“I think predicting the difficulty is really hard. Because you think it’s easy or it’s hard. But we’re not subject matter experts or students, and for subject matter experts, they are not students. So it’s really hard to know. Is it too easy or is it too hard for students? (P2)”*

Lack of support on question authoring

Participants did not mention specific tools to support them design questions or instructional activities. They would either revise materials from past years or open a word document and start from there. One instructor kept a list of multiple-choice questions they can import to different platforms. Multiple instructors mentioned using Gradescope for grading. One instructor said that the past exams of their course become unused. They wanted to have a technical solution where all past exam questions are stored and they could know from the system which questions were used in which years.

Collaboration between learning engineers and domain experts is critical

From the instructional designers’ perspectives, they all mentioned that the collaboration between them and the domain experts is critical at all stages of instructional design, including coming up with learning objectives, refining learning objectives, break down learning objectives to specific skills, design activities, design assessments, etc. Since experts have limited time, drafting materials first and have experts refine them was found to be helpful, and coming up with ways to present materials so domain experts can quickly go over them and offer feedback is important.

5.2 Iterative design, Co-design of QuizMaker

Over the past two years, I have applied UpGrade on 9 modules related to research methods in three courses at CMU, including User-Centered Research and Evaluation (7 modules), E-learning Design (1 module), and Applied Machine Learning (1 module). Five instructors have been involved in this iterative design and development process. My past implementations of UpGrade require support from an engineer in the workflow, as shown in Figure 5.1. And in the past implementations, I played the role of the engineer in this process. On the one hand, with learning engineer as an emerging profession, I envision my procedures can be replicated by someone else. On the other hand, through these implementations, I explore opportunities to enable more independent use of UpGrade by the instructor.

An example use of UpGrade includes these steps:

- Step 1: Engineer (E) gets past student data from the instructor.
- Step 2: E runs an algorithm to segment the data, and this step requires manual check and cleaning depending on the structure of the input data.
- Step 3: Step 3 happens through a 45min-1h meeting between E and the instructor. During the meeting, they see segmented student solutions from the past and tag dependency of components. Usually, within this time, the instructor can specify 4-5 question schemata.

Workflow of UpGrade

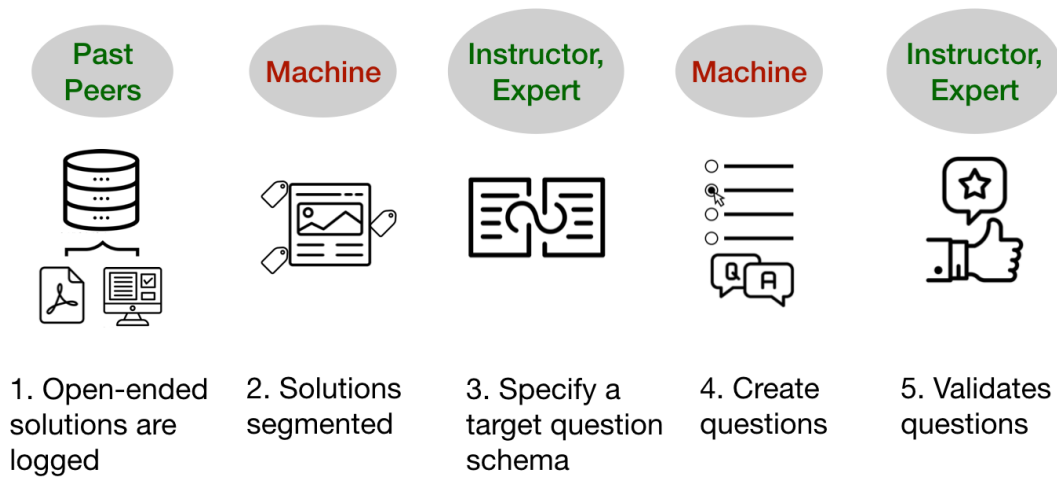


Figure 5.1: Five Step Workflow of UpGrade

- Step 4: E runs an algorithm using the specified data sources, a question pool is created
- Step 5: Step 5 also happens through a 1h meeting between E and instructor to select/revise questions. The usual yield rate is 20MCQs/45min-1hr.

In my past procedures, it takes 2 hours of instructors' time at most, and for the engineer, the solution segmentation in step 2 requires the most manual work. When the data source follows a certain template, or the data is logged hierarchically through Google forms, this step is much easier. Through my repeated observations, I also see opportunities to make the use of UpGrade more independent, and I'm developing a teacher interface through an iterative design process incorporating the instructors' feedback along the way.

After the instructor logs in to the interface and selects a module to import, they'll first look at past student solutions on this topic, Figure 5.2 shows a mapping between the original submission and the segmented solutions, which our users preferred to get a full picture of what student solutions look like.

The next step is to create question schemata, as shown in Figure ???. Instructors can select which components they'd like to display and specify the source component in the middle panel; an example question of the specified schema is displayed on the right. For example, the instructor can specify they want to use similar student answers as distractors, and the system computes the word vectors similarity between answers and selects the most similar ones as distractors. In the next step, instructors will select and revise questions. They will keep the ones that they like or modify some questions later. The iterative design process is still ongoing and I'll leave it to future work to fully evaluate the system.

QuizMaker Storyboard xuwang

View Assignment Create Question Group Create Question Group (A) Review Question Final Question Pool

PDF view Component view

C. Storyboards

Cover page (Storyboard 1-1)
User need: More personalized information on shuttle service
Lead question: Have you ever search for a shuttle route and stop near your home?
 Would you ever want to just search your destination and get the results?
Discussion points:

- Would you find this search feature convenient?
- How will you trust the search results?
- Will you feel that your privacy being violated since you are putting your address into the system?
- Would you want multiple search results in order to compare routes?

Shuttle Website Search Feature

Speed dating feedback:

- I will find this search feature convenient since right now when I need to find routes, I have to look at the routes individually and look at the maps to identify whether this route works for me.
- I would want to have more information on how does the system generate the suggestions and how long will it takes from school to the stop.
- I like this idea but I would like to have more suggested options to compare.
- Searching is nice, but I would like it not only for shuttle service but for all transportation service altogether, so that I can compare for example walking vs. shuttle or bus vs. shuttle.
- I kind of like the idea of searching, it is like google map but for shuttles.

Unit 1
Storyboard

1

Shuttle Website Search Feature

User Need for the Storyboard
More personalized information on shuttle service

Lead Question for the Storyboard
Have you ever search for a shuttle route and stop near your home? would you ever want to just search your destination and get the results?

Discussion Points
Would you find this search feature convenient? how will you trust the search results? will you feel that your privacy being violated since you are putting your address into the system? would you want multiple search results in order to compare routes?

Storyboard

2

Shuttle Stop Interactive Display

Previous Next

Figure 5.2: Screenshot of a prototype system: users see a mapping between the original submission and the segmented solutions.

- View Assignment
- Create Question Group
- Create Question Group (A)
- Review Question
- Final Question Pool

Storyboard

Answer 1

She's sitting on the bench. It's 8 pm and she's heading back home. She looks out of the window to see what she is at, but it's too dark and she doesn't know which stop the shuttle is approaching.

She takes out her phone and opens an app. The app tracks her location and tells her directly which stop she's approaching.

It turns out that the next stop is her destination stop. She quickly stands up and waits to see what other riders are going off. She gets off the shuttle and waves at her destination.

Target Question Stem:

- Textbox - Select Component

Which of the following user need is this storyboard designed for?

G

None

Storyboard

User Need for the Storyboard

Lead Question for the Storyboard

Which of the following user need is this storyboard designed for?

1

She's sitting on the bench. It's 8 pm and she's heading back home. She looks out of the window to see what she is at, but it's too dark and she doesn't know which stop the shuttle is approaching.

She takes out her phone and opens an app. The app tracks her location and tells her directly which stop she's approaching.

It turns out that the next stop is her destination stop. She quickly stands up and waits to see what other riders are going off. She gets off the shuttle and waves at her destination.

User Need for the Storyboard

Answer 1

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Answer 2

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Answer 3

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Answer 4

Riders need a reliable centralized channel with all shuttle information

Answer 5

Lead Question for the Storyboard

Target Correct Answer:

- Select Component

None

Storyboard

User Need for the Storyboard

Lead Question for the Storyboard

Target Incorrect Answer:

- Select Component - Criteria - Number

None

Storyboard

User Need for the Storyboard

Lead Question for the Storyboard

None

Random

Similar

Different

3

Riders want to know which stop they are approaching while on the shuttle in order to get off successfully at their destinations.

Riders want to know the waiting time for their shuttle in order to determine which transportation method to use. Similarity:0.6123724356957946

Riders need to know the current capacity of the buses to make better decisions. Similarity:0.3849001794597506

Having more confidence in how to ride the shuttle when boarding Similarity:0.307728727448319

Figure 5.3: Segmented past student work is displayed to the left; Instructors can create question schemata and preview the question on the right

Chapter 6

Using Psychometric Methods to Support Automatic Item Generation and Evaluation

Getting deliberate practice is critical in developing skill mastery. However, providing quality support to learners requires a tremendous amount of time and effort from instructors or other content providers. In college classrooms, for example, it is challenging for instructors to design sufficient high quality learning materials, including formative and summative assessments. Prior approaches have demonstrated the potential of automatically creating a large question pool through sourcing existing student open-ended solutions. However, challenges remain, such as selecting high quality question items from the pool. In this paper, we investigate the use of psychometric methods to prune out low quality items with small-size student performance data. We demonstrate in a classroom experiment that this question creation pipeline can produce high quality question items.

In this chapter, we investigate how crowdsourced data can be used to detect reliable versus unreliable question items. More specifically, we ask two research questions, *(i)* Can we use crowdsourcing to determine which items are unreliable? *(ii)* If so, how large a crowd is needed and how do we ensure the consistency of crowd workers?

6.1 Cronbach's Alpha to Evaluate Consistency

Cronbach's alpha [27] is a common psychometric measure of internal consistency across question items within a test. For a test, zero means no consistency at all whereas one indicates perfect consistency.

We use it to *(i)* evaluate the reliability of a set of UpGrade-created questions, and *(ii)* identify reliable and unreliable questions. To identify reliable and unreliable items, we first compute an overall Cronbach's alpha on a set of N questions. Then for each of the N questions, if Cronbach's alpha increases when the item is dropped, the question is indicated to be inconsistent with the rest of the questions, thus being a unreliable item, and vice versa.

6.2 MTurk Study

We conducted a validation study on Amazon Mechanical Turk to evaluate the quality of UpGrade-created question items. We focused the validation study on rubric item one in the heuristic evaluation assignment – identify heuristic problems from given scenarios. Figure 3.7 shows an example UpGrade-generated question on this rubric item. As shown in Table 3.4, 478 multiple-choice questions were created. We randomly selected 30 questions from the pool to evaluate their quality.

6.2.1 Participants and Procedure

We recruited participants from MTurk located in the US, with greater than 95% assignment approval rate, and more than 500 HITs accepted. Participants first spent 10 minutes reading about heuristic evaluation. Participants then proceeded to complete 30 multiple-choice questions about heuristic evaluation shown in a random order. The task took roughly half an hour to complete, resulting in an hourly pay of ~\$8/hour. A total of 70 participants completed the task. On average, participants spent 21 minutes answering all 30 questions, with an accuracy of 50%. To check whether crowd workers were randomly picking responses, we computed a user-user correlation matrix. Results show that among the $70 \times 69/2 = 2415$ participant pairs, all pairs had a correlation above 0.85, and 2405 (99.6%) pairs had a correlation greater than 0.9. This suggests that despite the low accuracy, participants were answering the questions carefully.

6.2.2 Prune Out Unreliable Question Items

The average Cronbach's alpha for the set of 30 items on the 70 participants dataset was 0.565. The correlation of each item with the total score, and the Cronbach's alpha after dropping this item are shown in Table 6.1. Using Cronbach's alpha as a criterion, 11 items were identified as unreliable items. Removing them resulted in a question bank of 19 items with a Cronbach's alpha of 0.74, suggesting high internal consistency in assessing student's heuristic problem identification. The 19 items were thus classified as reliable items.

6.2.3 Question Face Validity Inspection

We further performed a face value inspection analysis to understand what features resulted in unreliable question items. We summarized three reasons when a question is unreliable: *(i)* Multiple answers could be correct; *(ii)* There was a lack of description about the scenario, so students had to guess the original content creator's intention. This was consistent with our interview with students after the classroom experiment; and *(iii)* The original open-ended solution was of low quality, *e.g.*, there was misconception in the original solution, the writing was ambiguous. The face value inspection analysis offers insights on ways to improve question reliability.

6.2.4 Cost-effectiveness of Quality Control

With 70 crowd workers' performance data, we successfully identified 11 unreliable items in the 30-question sample. However, it may be unrealistic to recruit a large population of crowd workers to prune out unreliable question items for classroom use. With the collected MTurk dataset, we further investigated the minimal crowd size requirement for cost-effective quality control. We used the identified 19 reliable and 11 unreliable items as an approximation of ground truth. We then conducted experiments with varying crowd sizes from 5 to 70. We computed the accuracy for each experiment against the ground truth using the formula: $(\text{True Positives} + \text{True Negatives}) / \text{Number of Items}$. For each crowd size, we did 100 iterations of random sampling, and computed the average accuracy. The change of accuracy by crowd size is displayed in Figure 6.1. We can already do a decent job of differentiating reliable vs. unreliable items with a crowd size of 25 (accuracy = 0.8), and with a crowd size of 50, accuracy can reach 0.9.

Further, we investigated the crowd size requirement if the goal was to identify a subset of unreliable items. We ranked all 30 items that have been tested by their score correlation with the total score (Table 6.1), and used this as an approximation of the question quality ranking's ground truth. We then conducted experiments to investigate the crowd size requirement for identifying the X least reliable items in our sample. In the experiments we varied two variables, the crowd size, and the X least reliable items in the sample. For each combination of crowd size and X , we did 100 iterations of random sampling and computed the average accuracy on detecting the X least reliable items. Figure 6.2 shows the average accuracy for each experiment. When the goal was to detect the one least reliable item, we achieved an accuracy of 0.95 with only five students. When the goal was to detect the three least reliable items, we achieved an accuracy of 0.8 with

item	corr	alpha	item	corr	alpha
1	0.53	0.52	16	0.31	0.55
2	0.51	0.52	17	0.27	0.56
3	0.49	0.53	18	0.25	0.56
4	0.45	0.53	19	0.24	0.56
5	0.45	0.53	20	0.22	0.57
6	0.44	0.53	21	0.21	0.57
7	0.43	0.54	22	0.17	0.57
8	0.42	0.54	23	0.14	0.58
9	0.39	0.54	24	0.10	0.58
10	0.39	0.54	25	0.05	0.58
11	0.38	0.54	26	0.00	0.58
12	0.38	0.54	27	-0.03	0.57
13	0.36	0.55	28	-0.04	0.58
14	0.34	0.55	29	-0.08	0.60
15	0.32	0.55	30	-0.19	0.62

Table 6.1: The Pearson's correlation of each question item with the total score and the average Cronbach's alpha for the set when the item is dropped. Higher correlation and lower Cronbach's alpha indicates higher reliability.

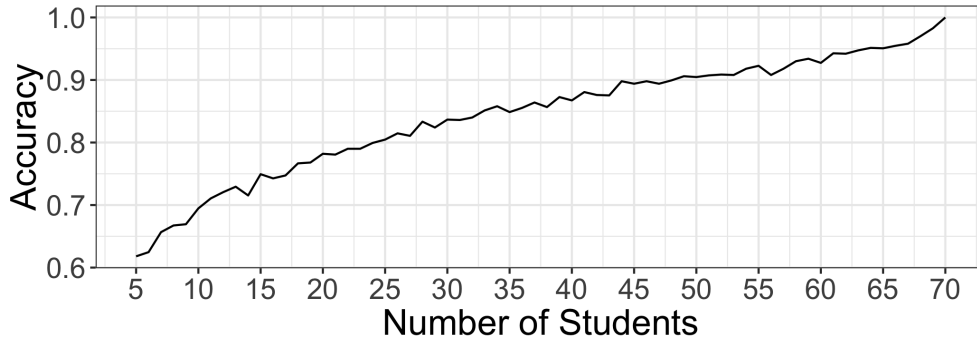


Figure 6.1: Average accuracy in detecting 19 reliable items and 11 unreliable items on different crowd size (across 100 iterations).

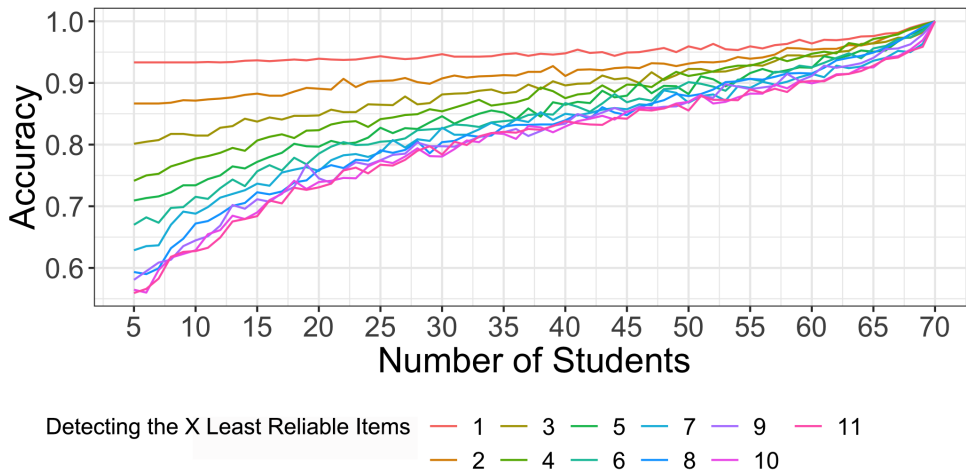


Figure 6.2: Accuracy in detecting the X least reliable items (X in the range of 1-11) varied by crowd sizes (across 100 iterations).

five students. These experiments demonstrated that more cost-effective quality control can be achieved depending on the needs.

6.3 Summary

Through quality control, we successfully identified 19 reliable multiple-choice questions that are highly consistent, with an average Cronbach’s alpha of 0.74. Considering the recommended reliability scores for exam use is 0.7-0.95 [130], the resulted question bank meets the criteria for classroom use. Proportionally, with the existing assignment data we have for UX101, we estimate UpGrade can output ~300 reliable multiple-choice questions on one rubric item after quality control. From a time consumption standpoint, if we hire crowd workers for quality control, assuming we prune out six questions in each 30-question set with 10 workers, UpGrade can generate 100 reliable questions with a minimal of 13 hours ($10 \times 4 \times 20$ minutes) of crowd workers’ time.

In contrast, it would take far more than 13 hours for instructors to write 100 reliable multiple-choice questions with feedback. Consider that the average time students spent to complete the open-ended assignment is 6.3 hours (as in the classroom experiment), which only includes five heuristic problems and corresponds to 5 multiple-choice questions. From a cost standpoint, it is nevertheless to mention it requires far more than $13 \times 8 = \$104$ to hire an expert to generate 100 practice questions. On the other hand, it might not be necessary to hire crowd workers for quality control. As more students use UpGrade, student performance data can be incorporated to prune out unreliable items, though with the risk of presenting low quality materials to students.

Chapter 7

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

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, in contrast to instructors' beliefs. 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. On the other hand, a study with 178 students in two college HCI classes 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 domains.

7.1 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, do scalable learning solutions come at the price of lower quality?

For example, online distribution of videotaped lectures is a powerful technique for scaling education but is in conflict with research suggesting that more interactive forms of learning-by-doing produce higher quality learning [29, 71, 110]. 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? In this paper, we address this question from both instructors' and students' perspectives.

MCQs are easier to grade especially with the availability of online learning and testing platforms, e.g., Canvas [1], GradeScope [120]. 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 [53, 59, 74, 113]. 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.

	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 7.1: 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.

Consistent with the quotes above, we first present evidence on college instructors’ beliefs about the relative value of MCQs and open-ended questions. 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 present evidence from student performance data that is surprisingly at odds with these beliefs. We designed 18 pairs of matched multiple-choice and open-ended questions on HCI research methods, including the topics of survey design and heuristic evaluation. 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. 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. In this paper we used question difficulty to indicate the potential learning benefit of the questions. This method follows theorized analyses and has been applied in the literature to help make instructional decisions [69]. We justify the method in the next section.

We suggest three general contributions of this work. First, our work indicates that, at least for some domains, 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 [98, 99]. Instructor beliefs are important because they will influence the design of curriculum and learning experience of students. In both this and a past case [69], 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 7.1. 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 [151]. 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.

7.2 Assessment Comparisons to Indicate Learning Benefits

Our ultimate goal is to explore the relative learning benefits (and costs) of the alternative question forms, i.e., multiple-choice and open-ended question forms. One low-cost path to that goal is examining the relative difficulty of both question forms through student performance data. In this paper, we used a question’s difficulty when used as assessments to indicate its potential learning benefit when used for instructional purposes. We describe the theorized analyses behind this argument as follows.

Open-ended questions that elicit constructed responses can be a kind of “desirable difficulty” [15], whereby students learn more by the constructive thinking processes than, for example, they would by simply reading text or watching lecture video. This desirable constructive thinking elicited by an open-ended question may not be required in a matched multiple-choice question. In this sensible line of reasoning, there are hypothesized cognitive demands in open-ended tasks

that are not present in a matched multiple-choice question. And the hypothesized extra cognitive demands in open-ended tasks are what make students learn more from them compared to matched multiple-choice tasks. We can use performance comparisons to test whether the hypothesized extra cognitive demands exist. If open-ended questions practice some constructive thinking processes that are not required in multiple-choice question answering, those open-ended questions should be harder for students to get correct than matched multiple-choice questions. Students who have not yet learned these important constructive thinking processes may nevertheless get the multiple-choice question correct, but get the open-ended question wrong. Thus, by this line of reasoning, the correctness rate on open-ended questions should be lower than the correctness rate on multiple-choice questions.

However, on the other hand, if we do not observe a difference in the correctness rate between matched multiple-choice and open-ended questions, this may suggest that the hypothesized extra cognitive demands required in open-ended tasks do not actually exist. In this alternative line of reasoning, i.e., there are no difference in the cognitive demands required in both tasks, both question forms would offer similar learning benefits to students. Following this theorized analyses, we designed an experiment using students' performance on matched open-ended and multiple-choice questions to infer the relative learning benefits of the two alternative question forms.

7.3 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 for learning. We discuss prior work that aimed at understanding instructor beliefs in correlation with their instructional actions. We finally 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.

7.3.1 Debate Around the Use of Open-ended vs MC 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., [111, 117]) 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 [128].

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 [37]. Surgue *et al.* found similar results in 7th and 8th grade physics class [128].

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 [123]. 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 [87, 88]. 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 [17]. 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.

7.3.2 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 [97]. 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 [98]. 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.

7.3.3 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 [128]. 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

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?" <i>A. What went wrong when you tried to open the file? Did it tell you it can't be opened because it expired?</i> <i>B. Did anything go wrong when you tried to open the file?</i> <i>C. What happened when you tried to open the file?</i>	2	Please indicate if the note is OK or describe(select) what is wrong with the interpretation note: "P1-35 Got a 404 error" <i>A. The note should refer to the participant</i> <i>B. The note is OK</i> <i>C. Not enough context is provided, e.g., why did they get the error</i> <i>D. It contains jargon "404" error</i>
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?" <i>A. No definition of "clear"</i> <i>B. Shouldn't ask users to suggest ideas</i> <i>C. Shouldn't ask about causality</i>	4	Please describe (select) the one most salient issue for the following interview question: "How often do you weigh yourself on a scale?" <i>A. Shouldn't ask a very personal question</i> <i>B. Shouldn't ask participants to estimate</i> <i>C. Shouldn't ask because of health care privacy laws</i>
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? <i>A. Recognition rather than recall</i> <i>B. Visibility of system status</i> <i>C. Aesthetic and minimalist design</i> <i>D. Flexibility and efficiency of use</i> <i>E. Match between system and the real world</i>	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"? <i>A. Display the "Register for a course" panel in the middle of the page</i> <i>B. Allow users to select courses and sections from a dropdown menu</i> <i>C. Only keep the "Search for a course to register" link on this page</i> <i>D. Display all available courses on this page</i> <i>E. Display help and documentation (e.g., steps to register for a course) in the middle of the page</i>
7	When conducting a think aloud, what would you do if a participant gets stuck on a task for 20 seconds? <i>A. Tell the participant how to move forward</i> <i>B. Wait longer before intervening</i> <i>C. Ask the participant to describe why they get stuck</i> <i>D. End the task early</i>	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." <i>A. "Sometimes" is vague</i> <i>B. It contains two different concepts</i> <i>C. "Students" is unnecessary</i> <i>D. The note is OK</i>
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?" <i>A. Which of these 2 smartphone cases would your colleague X like better?</i> <i>B. Which of these smartphone cases do you like the most?</i> <i>C. Can you rank these 3 smartphone cases' popularity among your colleagues?</i>	10	Please indicate if the note is OK or describe (select) what is wrong with the interpretation note: "P1-8 He is male" <i>A. The note is OK</i> <i>B. "He" is unnecessary</i> <i>C. This should go in the profile</i> <i>D. Not enough context is provided</i>

Figure 7.1: 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.

matched hands-on and paper-and-pencil tasks and showed consistent results [142]. Koedinger and Nathan designed matched algebraic problems in three formats, story problem, word equation and symbolic equation [69]. 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.

7.4 Methods

7.4.1 Hypotheses about Students' Underlying Cognitive Processes When Answering Questions

As introduced earlier, in this paper, we examined multiple-choice and open-ended questions' difficulty when used as assessments to infer their potential learning benefits when used as in-

structional activities. To form hypothesis about the relative difficulty of matched multiple-choice and open-ended questions, we performed a theorized analysis of the cognitive processes required when answering them.

There is an overlap between the cognitive processes required in answering MCQs and open-ended questions. If we break down the cognitive processes required in answering an open-ended question, it involves *(i)* generating candidate solutions and *(ii)* evaluating which solution is the best; In contrast, answering a multiple-choice question only requires *(ii)* evaluating candidate solutions. Thus, the important nuance here is how difficult it is to generate solutions compared to evaluating the candidate solutions. Consider the example of playing chess, generating a legal move might not be the real challenge; instead, the real challenge lies in evaluating which move puts the player in the best position. We hypothesize that for content domains where generating candidate solutions is especially simple and the real challenge is to evaluate which candidate solutions are incorrect or inadequate and which are correct or adequate, multiple-choice questions could be as hard as open-ended questions.

Empirically testing the relative difficulty of matched pairs of open-ended and multiple-choice questions can help us address this hypothesis. More specifically, if we found MCQs to be consistently much easier than matched open-ended questions, that would suggest that open-ended questions exercise more challenging generative processes (critical thinking skills) that are not exercised by MCQs. Thus MCQs are less valuable for probing deeply into a student's knowledge and skill; In contrast, if we found MCQs to be equally hard as or even harder than open-ended questions, that supports our hypothesis that there is an overlap between the cognitive process required in answering MCQs and open-ended questions. That would also suggest that for the tested domains, MCQs may exercise similar skills as their open-ended counterpart. In those cases, answering MCQs practice the challenging and critical underlying thinking elements about the content domain, just as or more so than open-ended questions.

7.4.2 Design of Matched Pairs of Questions

In order to assess the relative difficulty of multiple-choice and open-ended questions, we designed 18 pairs of multiple-choice and open-ended questions that have the same question stem, the only difference being the question format. The questions cover 4 topics around HCI research methods, 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. The questions are designed collaboratively with instructors who are teaching these topics at an R1 institution.

Example pairs of multiple-choice and open-ended questions used are shown in Figure 7.1. 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. To understand instructor beliefs about the difficulty of these questions, we first ask instructors to predict the relative difficulty of the pairs of questions. We then use the questions in actual HCI courses to get student performance data on them.

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

2. 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-choice	Open-ended	Same	Others (Please explain)
-----------------	------------	------	-------------------------

3. 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-choice	Open-ended	Same	Others (Please explain)
-----------------	------------	------	-------------------------

Figure 7.2: Survey questions that ask about instructors general belief about using MCQs and open-ended questions in their teaching.

7.5 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 instructor who are teaching university level HCI research methods courses. The survey is composed of two sections, the first section asks about instructors' general beliefs on using multiple-choice or open-ended questions in their teaching. The second section asks participants to predict the relative difficulty of pairs of questions.

7.5.1 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 section of the survey. 12 participants completed the second section of the survey.

7.5.2 General Beliefs

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. Some of the questions in section 1 of the survey are displayed in Figure 7.2. 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 2, 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 3, 45% of the instructors considered 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.

7.5.3 Predictions on Relative Difficulty of Matched Questions

The second section asks instructors to predict the relative difficulty of 10 pairs of multiple-choice and open-ended questions. We selected 10 pairs from the 18 pairs, as shown in Figure 7.1. 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.”

For 62% of the time, instructors answered that the open-ended question is harder than the multiple-choice question; 23% of the time, they answered that the multiple-choice question is harder than the open-ended question; For 15% of the time, they thought multiple-choice and the open-ended question are of the same difficulty.

7.5.4 Instructor Reasoning

In the survey, we also asked instructors’ views about the skills exercised by multiple-choice and open-ended questions respectively. Instructors tend to believe that MCQs mostly exercise recognition, and open-ended questions may exercise a larger variety of skills. Here are some instructors responses.

“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

“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

“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

Following the predictions of relative difficulty of the question pairs, we also asked participants to elaborate on their decision making process. Some explanations follow the contrast between recognition and recall as well.

“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

7.6 Classroom Experiments

This experiment is designed to address our hypothesis about the underlying cognitive processes involved in answering multiple-choice versus open-ended questions. This experiment also allows us to check alignment between instructor belief and student performance data.

7.6.1 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 an 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, 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.

7.6.2 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. Multiple-choice and open-ended questions were graded using the same rubric, 1 being correct and 0 being incorrect. 0.5 point were occasionally given (32

out of 1406 cases) to answers that addressed the intended problem but displayed additional errors. 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).

7.6.3 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 to build the model, and the formula is shown below:

$$score = question_form + (1|student_id) + (1 + question_form|question_id) \quad (7.1)$$

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 (7.2), 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.

$$\text{logit}(score) = (\beta + \beta_j) * question_form + \alpha + \alpha_j + \alpha_i \quad (7.2)$$

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.

7.6.4 Instructor Reasoning and Student Data Conflicts

We compared instructor prediction of the relative difficulty of matched multiple-choice and open-ended questions with student performance data. Table 7.2 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 (7.2). The odds ratio shows to what extent the open-ended format of the question is easier than the multiple-choice 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 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 7.2 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.

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 7.2: Ranks the 10 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	13	15	0	28
Same	6	12	0	18
OE	17	57	0	74
Total	36	84	0	120

Table 7.3: This table shows counts of observations where instructor predicted this question format to be harder and observations where student performance data suggests this question format to be harder. The greyed area shows when the two aligns.

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?"

- S1: Alternative (1pt):
 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?
-
- S2: Alternative (1pt):
 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?
- S3: What happened when you tried to open the file?
- S4: Could you ~~help~~ walk me through what happened when you tried to open the file?
- S5: What error message did you get when it ~~was~~ went wrong when you tried to open the file?

Figure 7.3: Example student answers in response to question 1 in Figure 7.1, including correct and incorrect answers for both formats.

7.6.5 Answer Examples

We show several example student answers in response to Q1 in Figure 7.1, including correct and incorrect answers for both formats. The example answers are shown in Figure 7.3. Some exam papers suggest evidence that students are evaluating and comparing options when they work on questions (S2). Often times, the wrong answer students give in open-ended questions assemble the incorrect options we present in the MCQs (S5). This also explains why MCQs are challenging and can be equally valuable as they target specific student misconceptions.

7.7 Discussion

Our experiment reveals that the contrast between multiple-choice and open-ended questions is not as simple as the contrast between recognition versus recall, especially when it involves learning objectives that are high on the hierarchy of Bloom's taxonomy, e.g., "Evaluation." For example, for Q3 shown in Figure 7.1 "Please describe the most salient issue for the following interview question", student performance data shows that the multiple-choice format is harder than the open-ended format. The challenging part in answering this question is not coming up with candidate solutions (potential issues for the interview question), rather it is evaluating whether each of the issues presented applies.

Our work demonstrates, for at least the domains we have tested, there is no evidence to support the hypothesis that MCQs are easier than matched open-ended questions. Indeed, we found cases where distractors were very competitive that made multiple-choice even harder. Competitive distractors often contain frequent student misconceptions, which require students to engage in challenging thinking processes that they may otherwise skip when working on an open-ended problem. Prior work showed that leveraging past student performance data can help instructors quickly create effective multiple-choice questions at scale [139]. Leveraging existing question creation methods such as [139, 147] can address instructors' potential concerns that multiple-choice questions are harder to design compared to open-ended ones as revealed from the instructor belief survey.

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. 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 7.1 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. The misalignment between instructor belief and student performance data further suggests that future instructional and assessment design should be theoretically and empirically rooted.

This work suggests that we need to establish the profession of Learning Experience (LX) designers to develop curriculum in higher education. Our work also demonstrates 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 [127]. 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.

7.8 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.

Chapter 8

Practice-Based Teacher Questioning Strategy Training with ELK: A Role-Playing Simulation for Eliciting Learner Knowledge

Practice is essential for learning. However, for many interpersonal skills, there often are not enough opportunities and venues for novices to repeatedly practice. Role-playing simulations offer a promising framework to advance practice-based professional training for complex communication skills, in fields such as teaching. In this work, we introduce ELK (Eliciting Learner Knowledge), a role-playing simulation system that helps K-12 teachers develop effective questioning strategies to elicit learners' prior knowledge. We evaluate ELK with 75 pre-service teachers through a mixed-method study. We find that teachers demonstrate a modest increase in effective questioning strategies and develop sympathy towards students after using ELK for 3 rounds. We implement a supplementary activity in ELK in which users evaluate transcripts generated from past role-play sessions. We demonstrate that evaluating conversation moves is as effective for learning as role-playing, while without requiring the presence of a partner. We contribute design implications for role-play systems for communication strategy training.

8.1 Introduction

With increasing challenges associated with the rapidly changing landscape of work, technology-based solutions that support professionals in lifelong learning are in more demand. Communication and interpersonal skills are critical for many professions, such as medical workers [58], researchers [19], and teachers [40]. One consistent challenge with the training of communication skills is that practice opportunities are limited and people are often expected to pick up the skills “on the job” [20, 49, 58, 153]. Role-play simulations show promises of providing practice and rehearsal opportunities for novices learning about communication strategies, for example, for medical students to rehearse doctor-patient conversations [58], for nursing students to develop nurse-to-doctor handover communicative competencies [20, 153], and for business students to learn and

apply influence tactics [49]. In this work, we focus on the training of questioning strategies for K-12 teachers and we discuss the implications of technology design for communication skill training across contexts.

Effective teacher learning and professional development is important for student achievement. In the United States, teacher professional learning programs provide insufficient opportunities for teachers to practice important skills and judgments [44]. “Pre-service” teacher training – the process of initial licensure in colleges of education – typically combines classroom-based course work with field observations and practicum teaching [81]. In-service teacher training programs take the form of professional seminars, workshops, expert or peer classroom observation and consultations [23, 57, 126]. For both pre- and in-service teachers, teacher learning consists primarily of attending lectures or discussions. Teachers have few opportunities to rehearse and practice teaching with students or receive meaningful feedback. Teachers are expected to gradually pick up skills “on the job” as they teach in real classrooms, in a haphazard process of trial and error with insufficient feedback. We align our work with a movement of teacher-educators advancing *practice-based teacher education*, an approach to teacher professional learning that emphasizes opportunities for rehearsal and reflection [154].

Although decades of research has shown that it is essential for teachers to understand what students know in order to help them learn [9, 25, 112], teacher practice is still far from satisfactory. Pre-service teachers often regard students’ misconceptions as barriers to learning, rather than useful starting points for instruction [79]. Even seasoned teachers often focus on evaluating student work as right or wrong, rather than understanding student work as evidence of current understandings that can be built upon [26]. This paper focuses on helping teachers develop the skill of *eliciting learner knowledge* through a role-play simulation system.

We introduce ELK (Eliciting Learner Knowledge), a role-playing simulation system that offers virtual sessions in which players can learn and practice discourse strategies on eliciting knowledge from their conversational partners. After two players join a role-play session, each of them assume the role of either a “Teacher” or a “Student”, and respectively receive teacher or student profiles. The two players chat through a text-based interface. The goal of the “Teacher” player is to elicit as much as possible of the “Student” player’s prior knowledge as indicated in their profile (Figure 8.1). Both players take a quiz in the end to assess and reflect on their performance.

We implement a supplementary activity in ELK –“Coding”, where players evaluate authentic transcripts generated from past role-play sessions. As shown in Figure 8.4, in this “Coding” activity, players assign a qualitative code to each line of the transcript, indicating which questioning move was employed – such as “Telling”, “Evaluating”, or “Probing.” The design of this feature is motivated by the constraints of role-play simulations and the demonstrated success of evaluation-type activities in other domains. For example, some role-play activities are found to be overly challenging and cause performance anxiety in players [49], organizers reported the overhead in pairing players when it requires multiple people to be present at the same time [24]. On the other hand, prior work has also demonstrated the strengths of evaluation-type activities for learning. For example, solving Parsons problems, i.e., evaluating correctness and ordering of coding snippets, is as effective for learning as writing the equivalent code[32], evaluating the quality of example survey questions is as effective for learning as writing new survey questions [139]. However, most of these prior work focuses on conceptual or technical skills that can be learned independently.

We explore the feasibility of employing evaluation-type activities for learning communication strategies, such as eliciting knowledge from a conversation partner.

We evaluate ELK with 75 pre-service teachers to answer three research questions. First, how effective is ELK in helping players develop questioning strategies? A mixed-methods approach was adopted in answering this question. To assess behavioral change, we look at how much improvement participants make in questioning moves from the first to the last (third) round of playing ELK. To assess conceptual and attitude change, we use an in-depth survey asking participants to reflect on their learning experiences with ELK. Second, does the “Coding” activity help participants develop questioning strategies? To address this question, we use the “Role-play” activity as the control condition, and compare their effects on helping users learn questioning moves. We also tease out the factor of receiving feedback from both conditions, resulting in a 2 by 2 experiment design, with the first factor being either in the “Role-play” or the “Coding” activity, and the second factor being whether users receive feedback or not. Third, what are users’ experiences in ELK and what challenges do they encounter? How can we better design computer-based role-play systems to make them more engaging and useful? We address this question through an open-ended survey sent to participants after they use ELK for at least an hour.

Our key contributions include:

- New system: ELK (Eliciting Learner Knowledge), a text-based role-playing system that enables pre-service teachers to practice questioning moves through simulated “teacher-student” conversations.
- Evidence of support for learning and sympathy development: An evaluation with 75 pre-service teachers show participants demonstrate a modest increase in effective questioning moves after using ELK for three rounds. Our study also provides evidence of conceptual and attitude change from players. This is the first study to our knowledge that conducts a thorough evaluation of the educational benefits of a computer-based role-play system, using both performance and subjective outcome measures.
- Benefits of the “Coding” activity: We show that evaluating authentic transcripts generated by others help participants develop questioning moves to a similar degree as generating improvisational questions in the role-play chat. The “Coding” activity has practical benefits as it can be performed by a single participant alone, which serves as a viable supplementary activity for online role-play systems.
- Design implications for role-play systems for communication strategy training: the evaluation study of ELK reveals prospects and challenges of role-play systems for communication strategy training. We summarize the design implications and discuss the broader application scenarios.

8.2 Related Work

In this section, we review prior teacher training programs and technologies that emphasize classroom discourse. We also discuss relevant literature around role-play systems for learning and simulation-based training. Finally, we present recent work on learnersourcing techniques. We discuss how these prior work motivated the design of ELK, and how our contribution situates in

these bodies of literature.

8.2.1 Teacher Classroom Discourse Training and Support

Teacher education researchers have explored a variety of approaches to improving discrete elements of practice such as questioning strategies. One method involves recording teachers' one-on-one conversations with students and have experts review them afterwards [122]. Another approach involves creating face to face simulations, where teacher educators act a students with misconceptions, and pre-service teachers practice questioning strategies. [119]. These are compelling approaches, but they are very demanding on teacher educator labor (to review video, act as students, provide individual feedback, etc.) We build upon these promising approaches by creating a digital platform for practicing question asking strategies in a peer-learning context, where teacher education students play both roles of teachers and students and provide each other feedback on discrete elements of practice: eliciting learner knowledge.

8.2.2 Role-play and Simulation Systems for Learning

Many other professions equip practitioners with extensive hands-on training and practice opportunity before they start their professional life, such as nurses [11, 96]. However, teachers often do not get sufficient opportunities to practice teaching in low stakes settings [44, 81]. When teachers do engage in low-stakes simulated practice, it is often in the form of “rehearsals” where participants practice teaching an entire activity or lesson to a group of simulated students [78]. Teaching is immensely complex, and research on complex learning suggests that novices often struggle to practice a whole complex assemblage while improving at specific elements of the task [66]. Rehearsals of the whole assemblage of teaching, therefore, should be complemented by opportunities to practice more discrete skills and judgments in teaching practice [109].

8.2.3 How to Elicit Learner Knowledge

Prior work has studied effective “talk moves” for teachers to elicit learner knowledge, such as asking follow-up questions [25, 31, 38, 95]. We reviewed prior work and then developed a framework for training pre-service teachers on effective questioning moves. The framework contains five categories of questioning moves.

- Priming: preparing the class for learning
- Eliciting: asking questions that reveal learner’s needs
- Probing: asking follow-up questions based on students’ responses
- Evaluating: responding in the positive or negative about students’ answers
- Telling: talking about the topic without listening to the student

Among the five questioning moves, we consider Priming, Eliciting and Probing to be effective ones in eliciting learner knowledge; and Evaluating and Telling to be ineffective ones for eliciting knowledge, since they are not optimal for teachers to understand what students know. (Telling and evaluating may be appropriate in other parts of the teaching sequence, but in the early phases of

eliciting learner knowledge, it is critical to simply *understand* student thinking before attempting to “fix” or redirect student thinking.

This five-part questioning framework is central to our pedagogical aims and research efforts with ELK. In implementing ELK in educational methods classes, teacher educators explain and demonstrate these questioning strategies to pre-service students. ELK participants then use these questioning strategies to code talk moves in transcripts. In analyzing ELK transcript data, as researchers we use this framework to identify pre-service teacher development – we consider an increase in the three effective questioning strategies and the decrease in the two ineffective strategies over multiple rounds of ELK to be evidence of pre-service teacher fluency in eliciting learner knowledge.

8.2.4 Learnersourcing and the Benefits of Solution Evaluation

We implement a supplementary “Coding” activity in ELK grounded in cognitive science and instructional design. As shown in Figure 8.4, users assign a questioning move to each line of an authentic transcript generated from previous role-play sessions. The design of this activity is motivated by the cognitive load theory [9, 34], aiming to reduce the cognitive load required in learners and allow them to focus on the learning task.

Prior work has shown that in some domains, evaluating the quality of solutions can support learning and performance on generating solutions afterwards, even with higher learning efficiency compared with practicing with generating solutions only. For example, Yannier *et al.* 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 [152]. Wang *et al.* 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 [139]. Ericsson *et al.* 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.

However, prior work mostly focused on technical skills that do not require interpersonal communication. For skills such as asking questions, it remains unknown whether evaluating responses can be a useful exercise, and whether it is more, less, or equally useful for learning as generating improvisational responses to scenarios. For ELK, the “Coding” activity would be especially helpful when a partner is not present, giving users more independence in using the system.

8.3 ELK: A Role-Playing Simulation System

ELK aims at helping teachers develop effective questioning strategies. The major function of ELK is a text-based role-play simulation, in which two players chat based on pre-written profiles. The goal is for the “teacher” player to develop effective questioning moves in eliciting the “student” player’s knowledge. We adopted a text-based interface, in which players type to communicate. Although this differs from the authentic teaching experience people may have, the goal of ELK is to provide focused practice for users to learn questioning moves without the

Algebra Grade 6

You are a 6th grade student in math class. Your teacher is about to start a lesson on variables and equations, and would like to see what you already know.

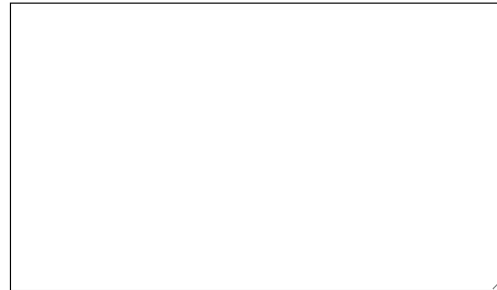
Your Student Profile:

Any letter given to me has the value of the where that letter is in the alphabet. A (or a) always has a value of 1, b always has a value of 2, and so on. Even if the problem tries to trick me by saying a=3, I will always input the value of 1 for a. I will always convert to numbers before I do anything else with the problem. If anything is given to me without operation symbols(+,*,-,/), I will just stick the number in front of that value. An example would be 3a. I know that a has to be 1. The value of this would be 31.

When you are ready to begin the round, click Begin

Begin

Use the scenario to the left to guide your conversation:



Chat here

Send Message

When the 7 minute round is finished, take the quiz.

Take Quiz

Figure 8.1: “Role-play” interface in ELK for the “Student” player. The profile is shown on the left, including the student’s (mis)conceptions about the topic. Players chat on the right.

ELK About Onboarding Scenarios Profile E-mail Sign Out

Algebra Grade 6 Participant Role: Teacher

Your Background
You are teaching a 6th grade class about variables and equations that involve variables. You would like to see what the students know before starting the lesson. No assignments have been given.

Your Objective
You would like to know how well your students solve math problems with very simple equations such as $x+p = q$, $px = q$ and $px+s=q$ for nonnegative rational numbers. This means that you both want to check their ability to compute simple expressions such as $2*a$ and solve equations.

ELK About Onboarding Scenarios Profile E-mail Sign Out

Algebra Grade 6 Participant Role: Student

You are a 6th grade student in math class. Your teacher is about to start a lesson on variables and equations, and would like to see what you already know.

Your Student Profile:

- Any letter given to me has the value of where that letter is in the alphabet. A (or a) always has a value of 1, b always has a value of 2, and so on. Even if the problem tries to trick me by saying a=3, I will always input the value of 1 for a. I will always convert the letter to numbers before I do anything else with the problem.
- If anything is given to me without operation symbols(+,*,-,/), I will just stick the number in front of that value. An example would be 3a. I know that a has to be 1. The value of this would be 31.

Figure 8.2: Teacher and student profiles on the topic of Grade 6 Algebra

cognitive load required in managing other aspects of their behaviors. ELK is publicly available at: <https://newelk.herokuapp.com>¹

¹Reviewers are welcome to use these test accounts to try out ELK. Pick one of the three usernames: [CSCWGuest1, CSCWGuest2, CSCWGuest3], the password is the same for all accounts: ELKGuest. Please note that for the “Role-play” activity, two players need to be present. For the “Coding” activity, please go to Onboarding and select “Learning to ELK as a teacher.”

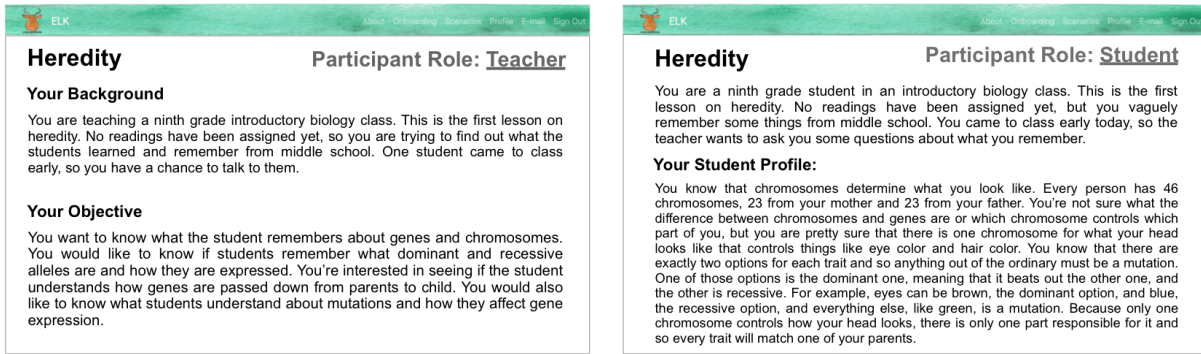


Figure 8.3: Teacher and student profiles on the topic of Heredity

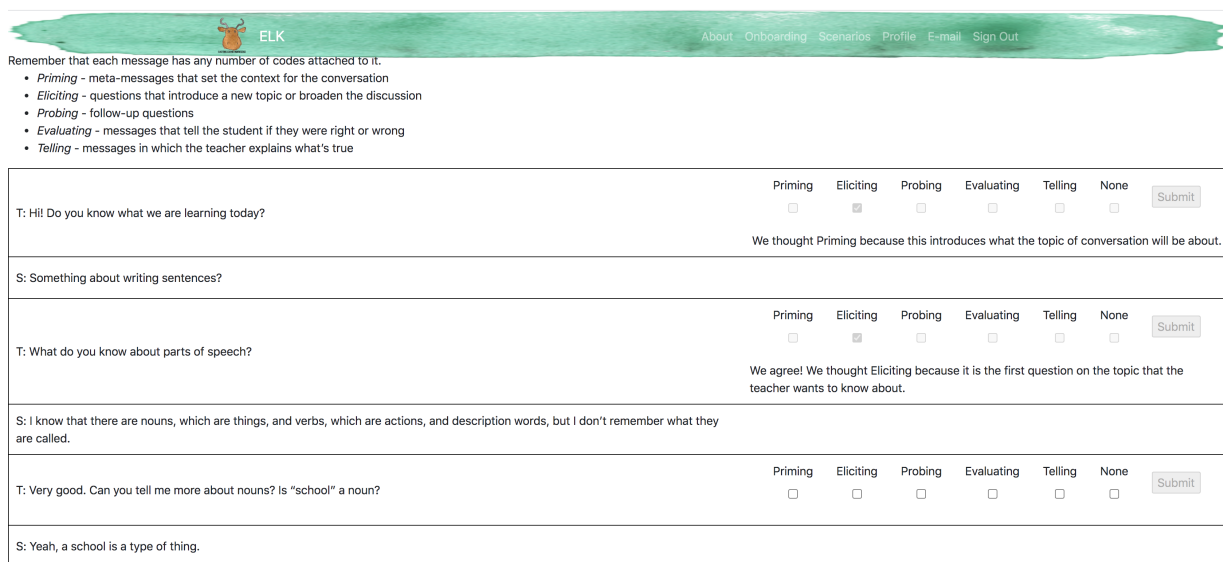


Figure 8.4: “Coding” activity in ELK. The user reads an authentic transcript generated from past role-play sessions and assign a questioning move to each line in the transcript.

8.3.1 Text-based Role-Play

Activity

The text-based role-play requires two players to be online at the same time and engage in conversations. Two players each take the role of a “student” or a “teacher”. After entering the platform, players first select a topic, e.g., grade 6 algebra, or grade 3 multiplication. The interface for the “student” role player is shown in Figure 8.1. The profile on the left specifies the prior knowledge held by the player, and the player should play out this persona. The “teacher” role player enters the same interface with a “teacher” profile. Two example pairs of “teacher” and “student” profiles are displayed in Figure 8.2 and Figure 8.3. These profiles are pre-written by

researchers in the area and senior K12 teachers. Players press begin to start the conversation, each session takes 7 minutes. The goal of the “teacher” player is to elicit as much prior knowledge from the “student” player as possible.

Feedback

Both players take a quiz in the end in which they answer three True/False questions about the (mis)conceptions in the “student” profile. For example, in the “student” profile shown in Figure 8.1, a misconception is that the variable a is always equal to 1. One corresponding question in the quiz is: Is this statement “ $a+4=6$ means that $a=2$ ” True or False? If the “teacher” player understood the “student” player’s misconception, they would answer False, otherwise True. The quiz gives the “teacher” player feedback on whether they successfully elicited the learner’s knowledge.

8.3.2 Coding Activity

Activity

In the “Coding” activity, users read authentic transcripts generated from previous role-play sessions. We apply the five-part questioning framework in the coding activity. Users first read descriptions about each of the five questioning moves and then assign a move to each line in the transcript, as shown in Figure 8.4.

Feedback

To enable real-time feedback to players, two experts from the development team coded 5 transcripts in the dataset. With the labeled transcripts, the system provides real-time feedback to users after they make a selection and click submit. Example feedback is displayed in Figure 8.4.

8.4 Evaluation of ELK

The evaluation study aims to address the following three research questions.

RQ1: How effective is ELK in helping users develop questioning strategies and understand student misconceptions?

RQ2: How effective is the “Coding” activity? How does it compare to the “Role-play” activity in helping participants learn questioning moves?

RQ3: What are users’ experiences with ELK and what challenges do they encounter? How can we better design systems for communication strategy training?

8.4.1 Participants

To address the above questions, we wanted to evaluate our system with pre-service teachers who are learning about classroom discourse and questioning moves. We reached out to teacher training programs at a small liberal arts college in the Mid-Atlantic region. Three faculties from the liberal

arts college signed up to use ELK in their classes. In total, 75 participants, who are enrolled in an undergraduate teacher education program, completed the study.

8.4.2 Study Components

In response to RQ1, we adopt a mixed-methods approach and use both performance and self-reported measures to assess participants' behavioral and conceptual change on eliciting learner knowledge. In order to examine participants' behavioral change on adopting effective questioning moves, we use a performance measure by quantifying the effective questioning moves "Teacher" players displayed in the role-play chats. We design the study so that participants use ELK for multiple rounds, which enables us to see their behavioral change over time. In order to gauge participants' conceptual and attitude change, we design an in-depth survey asking participants to reflect upon their experiences.

In response to RQ2, in order to understand whether the "Coding" activity is helpful, we design an experiment to compare the effectiveness of the "Coding" activity with the "Role-play" activity in helping participants adopt effective questioning moves. One consideration here is that the "Role-play" and the "Coding" activities have different feedback mechanisms. Feedback for "Role-play" is provided through a post-hoc quiz, and feedback for the "Coding" activity is provided immediately as participants evaluate transcripts. To make a fairer comparison and make the results applicable for cases when feedback is not readily available, we decide to tease the effect of feedback apart from the activity itself. We administer two versions of both activities, a version without feedback, and a complete version as introduced in Section 3.

In response to RQ3, we include questions in the survey asking about participants' experiences and the challenges they encounter when using ELK. In the study design, we make sure that all

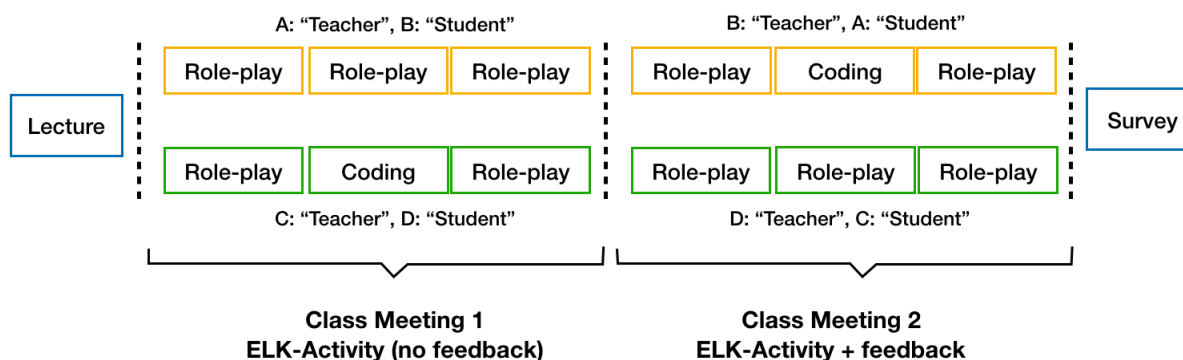


Figure 8.5: The study spans two 90-minute class meetings. In the first class meeting, participants play three rounds of ELK without feedback. This means for the "Role-play" activity, the quiz is disabled, and for the "Coding" activity, participants do not get feedback after making a selection. Participants are paired and assigned to one of the two conditions. The only difference is whether they complete the "Role-play" or the "Coding" activity in the second round. In the second class meeting, participants play three rounds of ELK with feedback. Participants remain in the same pair and switch roles. The pairs that did the "Coding" activity in Round 2 will now do the "Role-play" activity in Round 5 and vice versa.

participants experience both “Role-play” and “Coding” activities, and have played both “Teacher” and “Student” roles in the role-play sessions, so that they can comment on all aspects of the design of ELK.

8.4.3 Procedure

The study was conducted as a part of a teacher education course at the liberal arts college. We communicated the experimental procedure with the three course instructors that signed up, and they implemented the same procedure in their classes. The study spans across two 90-minute class meetings. An overview of the procedure is shown in Figure 8.5. Overall, there are four groups of participants, labeled as A, B, C, and D as shown in Figure 8.5. As an example, the sequence of activity for a participant in group A would be: Round 1 - “Teacher” in role-play with no feedback; Round 2 - “Teacher” in role-play with no feedback; Round 3 - “Teacher” in role-play with no feedback; Round 4 - “Student” in role-play with feedback; Round 5 - Coding with feedback; Round 6 - “Student” in role-play with feedback.

In the first class meeting, the instructor first gave a short lecture on how to elicit learner knowledge, covering effective and ineffective questioning moves. Participants then used ELK for three rounds, for about 30 minutes in total. Participants were randomly divided into two conditions and were paired within each condition. For each pair, one participant assumed the role of “Teacher”, and the other assumed the role of “Student” for the entire class meeting. The only difference between the two conditions is whether they did the “Role-play” or the “Coding” activity in the second round. For the “Role-play” sessions, the “Teacher” profile is consistent across three sessions to reduce the amount of background knowledge and reading required, but the “Student” profiles are different in all sessions to make sure the “Teacher” role still needs to adjust their questioning moves to be able to elicit knowledge from their partners. For the first class meeting, we disabled the feedback function for both activities, meaning participants do not do the quiz at the end of the “Role-play” chat, and do not receive feedback when they evaluate the transcripts.

In the second class meeting, all participants remained in the same pair and switched roles. This means that the participant who played the “Teacher” role will now play as a “Student” with the same partner. The pairs that did the “Role-play” activity in Round 2, e.g., A and B, will now complete the “Coding” activity in Round 5, and vice versa. Feedback is enabled for both the “Role-play” and the “Coding” activities.

At the end of the second class meeting, all participants would have experienced all roles and activities in ELK. We sent a survey via Google Form that asked participants to reflect on their experiences in ELK. We consider this design to best utilize the participant resources we have and provide us with insights about the effectiveness of ELK.

8.4.4 Outcome Measures

Frequency of Effective Questioning Moves in Role-play Chats

This learning outcome measure only concerns participants who play the role of “Teacher” in the role-play sessions. We developed a coding framework to gauge the quality of questions asked by

the “Teacher” player. The coding framework was developed based on prior work on how to elicit learner knowledge. As we develop the coding manual, we make sure that on the one hand this can cover the nuances between questioning moves we have seen in our pilot data, such as “Eliciting” and “Probing”. On the other hand, we also make sure this connects with existing work on what types of questioning moves are effective, so that it can help us quantify the quality of questions being asked by the players.

A brief version of the coding manual is shown in Table 8.1. The full version is provided as supplementary material for this submission. In the coding manual, for each category of questioning move, we provide multiple examples including explanation of edge cases. As shown in 8.1, “Priming”, “Eliciting” and “Probing” are considered to be effective moves in understanding what the students already knew, whereas “Telling” is considered as an ineffective move towards eliciting student knowledge and does not support student-centered classroom conversations. “Evaluating” moves often lead to “Telling” messages that also do not contribute to understanding student (mis)conceptions. When counting the frequency of effective questioning moves in a chat, we summarized the number of “Priming”, “Eliciting” and “Probing” messages uttered by the “Teacher” player. The five categories are not mutually exclusive, meaning each line of the teacher’s dialogue can have multiple codes. As shown in Figure 8.7, the orange texts indicate the questioning moves assigned to each line of the “Teacher” role’s dialogue. The excerpt to the left has a total of 5 effective questioning moves, and the excerpt to the right has a total of 9 effective questioning moves. The coding framework enables us to quantify the quality of questions asked by the participants.

The coding manual was developed and refined through an iterative process. Two of the authors first independently coded subsets of the dataset and addressed disagreements, enriched the definitions and examples in the coding manual, and added categorization for edge cases. When the coding manual is complete, two of the authors independently coded 100 lines of “Teacher” dialogues and reached high agreement on all 5 categories in the coding manual, with Cohen’s Kappa of 0.73 for “Priming”, 0.8 for “Eliciting”, 0.72 for “Probing”, 0.76 for “Evaluating” and 0.78 for “Telling”. One author continued to code all transcripts in our dataset following the coding manual.

Open-ended Survey

In addition to behavioral changes, we are also interested in knowing participants’ experiences of using ELK. To achieve this goal, we administered an open-ended survey at the end of the second class meeting. The survey questions include:

- *Please let us know more about how ELK may or may not have helped you learn questioning strategies for eliciting learner knowledge*
- *Please let us know more about how ELK may or may not have helped you learn about students’ conceptions about a specific topic (e.g., algebra or heredity)*
- *Did you change the way you asked questions during the game? Please elaborate on your answer above.*
- *What, if anything, did you find difficult about role-playing in ELK?*
- *What suggestions do you have for improvements to ELK?*

Questioning Move	Definition	Example
Priming	These are meta-messages that set the context for the conversation. They might appear at the beginning of the conversation or later to bring the conversation back to the topic/goal.	[after a student asks, “Is that right?”] I will tell you all about that during class today, but for now I just want to understand your ideas.
Eliciting	These are questions that introduce a new topic or broaden the discussion.	Teacher: If I were to give you $x+p=q$, would you know how to solve? What do you know about the area of a circle?
Probing	These are follow-up questions that go deeper into what the student thinks. It is often impossible to tell the difference between Eliciting and Probing messages without context.	Teacher: What do you know about word order? (Eliciting) Student: The noun comes before the verb. Teacher: Is this always true? Or are there ever nouns after verbs in a sentence? (Probing)
Evaluating	These messages tell the student if they were right or wrong, either explicitly or implicitly. They often lead to Telling messages and distract the student from the goal of figuring out their preconceptions.	That’s right. Not quite.
Telling	These are messages in which the teacher explains what’s true. While important during instruction, they are distracting if the goal is to figure out what the student already knew or believed.	The circumference of a circle is $2\pi r$.

Table 8.1: Brief version of the coding manual, with definitions and examples of the 5 questioning moves.

8.4.5 Survey Analysis Method

After we see the survey responses, we realized that many participants shared their experiences, takeaways, and challenges in their answers regardless of which question they were responding to. In our analysis, we broke down the boundaries between the questions and conducted a thematic analysis [18] of all the participants’ responses. First of all, two of the authors read and familiarized themselves with all the responses. They then did open coding of all of the responses independently. The two authors met and went over each of their comment and merged all the ideas into a list of themes. The two authors discussed and summarized 6 major themes from the data with a list of sub-topics within each theme. We will present our findings in response to each of the research question in the next section.

Participants #	Round 1	Round 2	Round 3	Feedback
19	Role-play	Coding	Role-play	disabled
18	Role-play	Role-play	Role-play	disabled
19	Role-play	Coding	Role-play	enabled
19	Role-play	Role-play	Role-play	enabled

Table 8.2: Distribution of participants across conditions

8.5 Results

In the experiment, all 75 participants played the “Teacher” role. The number of participants in each condition is shown in Table 8.2. Through a thematic analysis of the survey responses, we identified 6 major themes, including 1) Awareness about eliciting learner knowledge, 2) Adaptation in questioning moves throughout the process, 3) Gaining perspectives about the student stakeholder, 4) “Coding” helps 5) Unfamiliarity with the content domain makes it difficult, 6) Tweaks about ELK features can make it better. We present both experimental results and survey finding in response to each of the three research questions.

8.5.1 Effectiveness of ELK

Participants displayed modest increase in effective questioning moves from Round 1 to 3

We ran a paired t-test on the total number of effective questioning moves from Round 1 to Round 3 for all participants across conditions. We see that participants displayed significantly more effective moves in Round 3 compared to Round 1 ($p = 0.01$). The average number of effective moves per chat in Round 1 was 4.9, and the average number in Round 3 was 5.6. The average number of effective moves across conditions is shown in Figure 8.6. We see a trend that for all conditions, participants show modest increase in positive questioning moves.

Here is one example of behavioral change from Round 1 to Round 3 by P68 (Figure 8.7). In Round 1, as shown in the transcript to the left, the participant used multiple “Telling” messages. P68 directly told the student what was the correct answer, without trying to understand why the student came to a wrong answer in the first place. However in Round 3, as shown in the transcript to the right, under a similar circumstance where the student made a mistake, P68 started to use “Eliciting” and “Probing” questioning moves to understand why the student made the mistake. We have attached two additional full transcripts from the study in the Appendix.

Participants showed more awareness about eliciting learner knowledge

One emerging theme from the survey responses is that participants disclosed they realized the importance of understanding student knowledge and began to set expectations about what they would experience when they became teachers. On the one hand, some participants mentioned that they became to realize how difficult it is to understand student thinking and the frustration they could experience as teachers, *“This just showed me how difficult it is to find deficiencies or*

misconceptions. You almost need to know what you are looking for. And open ended question will not suffice sometimes. I enjoyed practicing as the teacher. It helped. (P37)” , “It gives me an insight as to what it is like to be a teacher and have students not understand the concept of a topic or not remember. It is very frustrating. (P21)”

Participants also shared the conceptual knowledge they have gained throughout the process, for example P42 said : *“It helped me to really know the difference between eliciting and probing and how to ask good, open-ended probing questions.”*, and P71 said : *“ELK helped me learn how to ask particular questions to get students to reconsider their answers and think of new ideas. Directly telling a student if they are right or wrong is ineffective and should be used sparingly. Teachers should guide students to new knowledge by asking questions that prompt their thinking.”*

Participants talked about adaptation of questioning moves throughout the process

This is an important theme we summarized from the survey responses. Participants talked about how they changed their tactics in asking questions as they engaged with ELK. Participants mentioned that they are now trying to ask more questions and explain less. *“I made a few changes to the way I asked questions and I ended up asking more questions rather than trying to explain the material (P44)”* More specifically, P8 who played the teacher role teaching the topic of heredity said: *“Instead of saying what a gene was and then asking a question I started from the very beginning I asked do you know what a gene is?”*

People also talked about switching from fact-based questions to more open-ended ones. *“Rather than asking yes or no questions, i tried to ask questions that made the student participate in the discussion. Open-ended, leading questions helped to let me know what the student knew.*

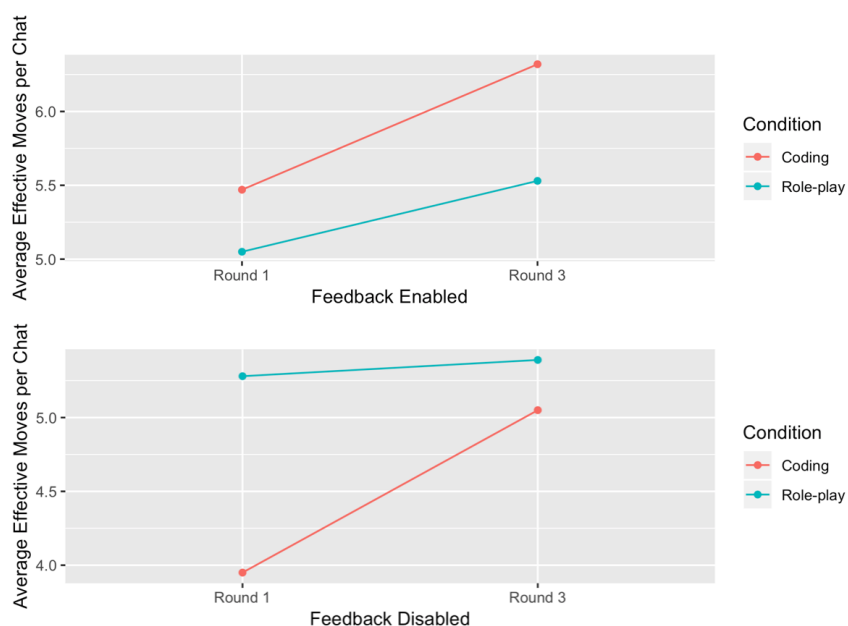


Figure 8.6: Participants increased positive questioning moves from Round 1 to Round 3 across conditions.

The figure consists of two side-by-side screenshots of the ELK (Educational Learning Kit) interface. Each screenshot shows a chat window with a teacher (represented by a red flag icon) and a student (represented by a graduation cap icon). The teacher's messages are on the left, and the student's responses are on the right. The interface includes a top navigation bar with 'ELK' and links for 'About', 'Onboarding', 'Sessions', 'Profile', 'Email', and 'Sign Out'.

Left Screenshot (Round 1):

- Teacher: "Okay, let's begin! Today we're going to be getting started on equations and variables within equations. What do you think the word 'variable' means?" (Priming)
- Student: "Variables are the letters in the equation. they stand for any number that solves the equation." (Priming, Eliciting)
- Teacher: "That's right! So if I asked you to solve $2a-4=4$, what do you think a would stand for?" (Evaluating, Eliciting)
- Student: "1" (Evaluating, Telling)
- Teacher: "If it were 1, that would make the equation $2(1)-4$, which would be $2-4$, which would then be -2 . So, we want something that will get us to the answer 4. How do you think we could solve for a?" (Evaluating, Telling)
- Student: "substitute a number for 'a'" (Telling)
- Teacher: "That's our end goal, yes. What number do you think makes the most sense to substitute?" (Evaluating, Probing)
- Student: "We want to end up back at 4, remember." (Telling)

Right Screenshot (Round 3):

- Teacher: "Welcome class, how are you all today?!" (Priming)
- Student: "Mehh, my brain hurts" (Priming, Eliciting)
- Teacher: "Oh boy, had a busy day, huh? Well, hopefully this won't be too hard! We're going to be going over equations and variables today, have you heard of those?" (Priming, Eliciting)
- Student: "Sv" (Eliciting)
- Teacher: "Good! So, what do you think they mean? What's an equation, and what's a variable?" (Eliciting)
- Student: "equation is a math problem that contains an =. A variable is a letter that is placed with in the equation." (Eliciting)
- Teacher: "Why is the variable there? What purpose does it serve?" (Probing)
- Student: "The variable is there to help solve the equation and show that there are many ways to solve it." (Probing)
- Teacher: "So, what you're saying is that the variable is a tool? Does it tell you what operation to use when solving an equation like $2a=8$?" (Probing)
- Student: "Sv" (Probing)
- Teacher: "Okay, so if I ask you to solve that equation, how would you do it? What does a tell you here?" (Probing)
- Student: "21 doesn't = 8 so the equation is not valid" (Probing)
- Teacher: "Where do you get 21 from?" (Probing)
- Student: "a=1 so naturally the number 21" (Probing)
- Teacher: "Why do you think that a=1? How did you come to that conclusion?" (Probing)

Figure 8.7: Left: excerpts from Round 1 of P68 playing the “Teacher” role. The participant used multiple “Telling” moves and directly told the student the correct answer when the student made a mistake. Right: excerpts from Round 3 of P68 playing “Teacher” role. When the student gave an unexpected answer, P68 used “Eliciting” and “Probing” moves to understand the thinking process of the student.

(P6)”, “Also asking questions that get the students to expand on what they know and why they knew what they did and how they got it. (P23)”

Participants mentioned adapting their questions to the student responses allowed them to more quickly get at students’ prior knowledge and misconceptions. For example, P20 mentioned how they changed from Round 1 to Round 2: “I started out asking standard questions about the topic. I made slight changes to my approach during the next round. In the 2nd round, I asked the student to explain their reasoning to me so that I could understand their misconceptions.” P50 and P8 shared their experiences of tweaking questions based on the students’ responses in order to more effectively elicit their students’ prior knowledge. “If the student seemed to know more about a certain area I would continue on that path. If it seemed like they didn’t I would veer to another realm of the subject. I tried to learn as much about what the student knew. (P50)”, “Some questions I asked only got me generic answers that, while still helpful, were only helpful in finding out what my student definitively knows about the basics, but not what they know about the subject at hand. So I was able to tweak my questions to get to the heart of the matter more quickly.”

Some participants referred to the questioning moves they have learned in ELK and shared

their experiences of using them. P8 and P29 talked about using probing and eliciting questions in their conversations: *“Probing questions, giving praise and minor corrections when needed to encourage the student to keep talking to me. (P8)”* *“I tried to ask a variety of general (eliciting) questions along with probing questions to get a good balance in the conversation (P29)”* While P49 shared that by changing the wording of the questions, the questions could be more useful in eliciting the learner’s prior knowledge. *“I really only changed the wording of my questions. Probing questions were still probing, but they became more like leading questions, where I was actively trying to get information out of the student, instead of just passively seeing what the student knows. (P49)”*

Participants disclosed that they gained the perspectives of students through playing ELK

This is an important emerging theme from the survey responses. Many participants talked about the benefit of switching perspectives in the role-play, and disclosed that knowing what a student may think through acting out as a student was “eye-opening”. Participants also talked about their willingness to be more patient with students when they become teachers in the future. For example, P8 said *“Students can come up with some odd associations, and seeing some of them written out helped remind me to be more flexible as a teacher because sometimes the associations make sense to others as well, and sometimes they only barely make sense to the student, so the teacher needs to be patient and open-minded.”* P17 gave a more concrete example of a student misconception they would never have thought of: *“When learning about variables, students thought the alphabet and the certain letter was paired up with where the number is in the alphabet. I never thought about it like that until then.”*

Participants also mentioned that playing ELK helped them understand that every student is different and thus understanding student thinking is critical yet challenging. *“It shows that all students have different knowledge and different ways of thinking, and teachers must adapt quickly to answer questions that their students have. (P74)”*, , *“ELK taught me that each student has a different concept of a topic. It is the job of a teacher to work through that and help the individual but also the whole class. (P50)”* , *“I learned that unless you ask the right type of questions students aren’t going to be able to explain what they know randomly. (P4)”*

8.5.2 “Coding” is an Effective Supplementary Activity in ELK

In order to evaluate the effectiveness of the “Coding” activity, we used the “Role-play” activity as a control condition in the experiment design. On the one hand, the feedback mechanism for both activities are very different. On the other hand, considering broader use cases of evaluation-type activities for learning communication strategies, expert annotated transcripts and feedback may not always be available. Because of these reasons, in the first half of the experiment, we disabled the feedback feature in ELK. This is to investigate to what extent is evaluating transcript helpful even when feedback is absent, when comparing to generating question moves in role-play sessions. In the second half of the experiment, we enabled the feedback feature in ELK, the goal is to compare both activities using the full setup in ELK.

In both cases, the second Round, being either “Coding” or “Role-play” is the intervention, and we compare their impacts on participants’ use of questioning moves from Round 1 to Round

	No Feedback Setting		With Feedback Setting	
	Coefficient Estimate	p-value	Coefficient Estimate	p-value
Condition (Role-play)	0.11	0.10	-0.04	0.65
Round (Round 3)	0.09	0.018*	0.07	0.13
Condition (Role-play)*	-0.08	0.13	-0.03	0.63
Round (Round 3)				

Table 8.3: Parameter estimates and p-value for both repeated-measures linear regression models comparing the effectiveness of “Coding” and “Role-play” activities. Both models show that there is no observed difference between the two conditions.

3. We run two separate comparisons for the first half (ELK without feedback) and second half (ELK with feedback) of the experiment.

We run a repeated-measures linear regression model on the long table format of the data, with each row being an observation. Each participant has two rows, with one being the performance for Round 1 and the other for Round 3. We use the number of effective questioning moves as the dependent variable, and the independent variables include Condition (a binary variable indicating whether Round 2 is “Role-play” or “Coding”), Round (a binary variable indicating whether this observation is from Round 1 or Round 3), and the interaction between Condition and Round. We included a random intercept for each participant ID in the model to account for individual differences. For both settings, the estimates of the parameters in the model are displayed in Table 8.3.

Both models show that participants demonstrated modest increase in questioning moves from Round 1 to Round 3 regardless of whether they did “Role-play” or “Coding” in Round 2. As shown in Figure 8.6, one big limitation of this experimental study is that the randomization did not work. In the no feedback setting, participants in the two conditions had significant difference in their performance in Round 1, with participants in the “Coding” condition being lower performing. Even though the “Coding” condition had a bigger leap in the no feedback setting, also indicated by a weak interaction between Round and Condition in Table 8.3, it is not clear whether it is because it is harder for participants in the “Role-play” condition to improve.

We want to point out here that having participants evaluate transcripts could be a useful activity for beginners. As shown in the comparison, this evaluation-type activity could help participants adopt effective questioning moves to a similar degree as the role-play sessions. It can be especially helpful when it is hard to find partners to role-play.

Coding activity experiences

In the survey, participants also talked about their experiences with the coding activity. Some participants found the “Coding” activity to be helpful for them to understand the questioning moves. *“Coding transcripts really helped me to understand the difference in question types (P37)”*, *“I liked being the teacher because I thought it was easier to be the teacher, especially after*

doing the coding. It was easier to guide the conversation and focus on gathering the student's knowledge. (P11)" , "It really had me thinking and what I was doing as a Teacher and a Student. I really liked the Coding activity we did, it helped out understanding the different ways teachers interact with the students."

Some participants said that the coding practice made them change their questioning moves. For example, P74 said: *"After the coding and switching roles I changed my answers by using the different types of questions and my questions also changed based on the students prior knowledge about the subject."*

8.5.3 User Experiences, Challenges and Feedback

Unfamiliar content makes it very difficult

This is the most prevalent theme we have identified from the survey responses. Many participants found the content domain to be critical for successful role-play. Some participants said that if they were not familiar with the content, e.g., heredity, it was hard for them to think of questions to elicit learner knowledge. *"Make the content easier and maybe geared toward a younger age group. It's hard to be the teacher when you don't know whether the student is right or wrong." , "Please provide teachers with a brief overview of the material because the content was really getting in the way. Or revise the material and make it simpler. "*

Trying to stay in character is hard

When taking the role of a "Student", participants find it hard to stay in character, especially when they do not have the same conceptions about the subject matter. Participants mentioned that trying to separate the role-play from their real-world knowledge is challenging. For example, P43, P2 and P3 said: *"Trying to stay in character was difficult. It was hard not to elaborate on answers with too much information before the teacher asked the question. You really had to have balance in how much you said and let the teacher try to pull it from you, rather than trying to help your friend who was playing the teacher." , "It's hard to separate my real-world knowledge from the student's. Acting as though I am confused or have limited knowledge is far harder than acting like I know more than I do. (P2)" , "it was kind of difficult because when I knew the answer was wrong I wanted to be able to answer it correctly and explain myself but I still had to follow the profile (P3)"*

Multiple opportunities and realistic experience through simulation

Participants enjoyed having multiple opportunities to tweak questions and appreciated the seemingly real experience provided by ELK. For example, P 42 said: *"It allowed me to play around with questions in order to elicit what they students already know (their conceptions).(P42)"*

P43 and P1 talked about having realistic experiences in ELK: *"ELK was really cool because it gave you seemingly real experience in trying to see what a student may or may not know about a topic". (P43)" "The student profiles seemed to have much more detailed information about what students knew and what they misunderstood. They seemed pretty realistic. I'm never going*

to be teaching algebra, but I can see how a student would be very confused and just think that $a=1$ no matter what. (P1)”

Provide more flexibility in the system

Participants also wanted to have more flexibility with the length of the sessions, with some hoping the sessions to be longer, and other hoping them to be shorter. *“Bit longer time for the rounds, please. Seven minutes is not really enough when you’re not talking audibly, and have to type; not everyone can type super fast. ” (P8) , “I felt rushed to ask so many questions in the short 7 minutes. I felt like time was up just when I was getting started.” (P59) “Maybe a 5 minute simulation instead of 7, and give more background for the teacher. ” (P56)*

8.6 Design Implications

From the above analyses, we summarize the design implications for role-play systems for communication strategy training.

8.6.1 Content Domain Knowledge is Essential

As mentioned by many participants, if they were not familiar enough with the content domain, it gets in the way of practicing communication skills. Participants also talked about providing a bigger pool of “Teacher” and “Student” profiles for players to choose from. For example, P20 said: *Please add different subjects. I’m a history and political science major. I don’t remember much from biology class. ,* and P1 said: *“Please include more topics that have to do with the humanities! I know very, very little at this point about algebra and biology.”* We are implementing a new feature in the system now that enables players to contribute “Teacher” and “Student” profiles. The goal of this is to enrich the profile pool to give players more options to choose from.

In addition to supporting teachers develop classroom discourse, this also applies to exercising other types of communication skills through role-play simulations. Allowing users to customize user profiles and providing background content knowledge to users could better prepare them to focus on developing the communication strategies of interest.

8.6.2 Role-Play Helps Participants Gain Perspectives of Relevant Stakeholders

In our study, many participants expressed that ELK made them aware of student misconceptions and the very different thinking processes students may have. Participants found acting out as a student is “eye-opening” for them to gain perspectives as a student. In addition to teachers, many professions are interpersonal and involve multiple stakeholders, such as doctors and patients, judges and victims, and mentors and mentees. When supporting these professionals develop communication competencies, role-play systems have great potentials and an adds-on benefit of helping learners gain perspectives from other relevant stakeholders.

8.6.3 Focused Practice through Evaluation Activities

The experiment suggests that having novices evaluate past transcripts can be an effective instructional activity, which also has the practical benefit as it can be performed by a single participant alone. For future communication strategy training programs, such evaluation-type activities can be applied before role-play sessions to foster understanding.

8.6.4 Different Modalities of Role-Play

ELK is a text-based platform where participants type to communicate. The design consideration is to reduce the cognitive load from players to manage their behaviors. In the survey, participants also mentioned the inconvenience of a text interface where they have to type. For example, P9 mentioned that “*maybe more of a face to face interaction, it take a lot of time to type the responses*” Future design of such system could provide different modes of interaction, e.g., as users become better in eliciting learner knowledge with the original text interface, they can go ahead and try a speech-based version.

One participant raised a point that it was hard for adults to imitate 6th graders in the conversation. “*It was hard to find a balance between talking as if I really was in 6th grade and saying things I would actually say as an adult (using words like "correspond," for example.*” Although the role-play sessions did not require the players to act like 6th graders, it is an interesting idea to scaffold players to talk in the same way as the persona specified in the profiles. This would offer a more realistic role-play experience for all participants.

8.7 Conclusion

All across the helping professions – teaching, medicine, social work, clerical work, and so forth – eliciting thoughts, feelings, and understandings from people is a critical part of professional practice. In this work, we demonstrate a system that supports simulated practice of questioning strategies through two learning modalities: question generation and question evaluation. In a teaching context, we find evidence that ELK helps participants value learner knowledge, empathize with the challenges of students as they develop understanding of STEM topics, and increase their use of questioning strategies that effectively elicit learner knowledge. We have tentative evidence that a combination of question generation and evaluation practice may be more effective than question generation practice alone in increasing the use of teacher questioning strategies.

While the scenarios in ELK are customized primarily to deal with STEM topics; ELK could be customized to support the development of interpersonal skills in a wide variety of contexts and professions. ELK’s digital platform supports a variety of educational implementations (in face to face classes, in online classes, as out of classwork, etc.) and collects data for participants, educators, and researchers to better understand how learners develop effective questioning strategies.

8.8 Appendix

Round 1 chat	Round 3 chat
<p>T: Good morning! Today we are going to be talking about Chromosomes and Alleles</p> <p>S: yippee</p> <p>T: Can you tell me about either of those? what do you already know about chromosomes</p> <p>S: I have 23 chromosomes!</p> <p>S: I got one chromosome from my mom and one from my dad and the last one is from both of them combined</p> <p>T: where did you get 23 from? you aren't too far off</p> <p>T: Yes that is true however you are missing some information S: Well, I have dominant and recessive alleles</p> <p>T: Well you aren't wrong you do have at least 23 chromosomes T: And you do have dominant and recessive traits!</p> <p>S: The extra chromosome I have is a combination of both T: You don't quite have an extra set of chromosomes, everyone has 46, did you think that you had 23 because you got a set from each parent?</p> <p>S: No? I have 23.</p> <p>S: One from each parent and the 23rd one is a combination from both of them</p> <p>T: You are correct you have at least 23 but the correct number is 46 you receive 23 from each parent</p> <p>S: I was told i have 23.</p> <p>T: I'm sorry but today we will learn why you have 46 chromosomes, it's going to be fun and we are going to learn a lot!</p> <p>T: What do you know about alleles</p>	<p>T: Good morning Class today we will be talking about genes, and chromosomes, and dominant and recessive traits!</p> <p>S: yAY</p> <p>T: Let's start with chromosomes! tell me what you understand about chromosomes</p> <p>S: Everyone has 46!</p> <p>S: 23 from my mother and 23 from my father</p> <p>T: YES that's correct! excellent! now on to dominant and recessive traits</p> <p>T: what do you know about these</p> <p>S: What are alleles?</p> <p>T: Do you know anything about them at all?</p> <p>T: Tell me what you know and understand first then we can move on from there</p> <p>S: Well i heard someone mention alleles earlier but I thought dominant and recessive traits were about chromosomes</p> <p>S: Because chromosome are like genes and there is one for each part of your body</p> <p>T: Alleles is just another term for gene or chromosomes, they all relate to heredity and traits</p> <p>T: Do you know what a recessive allele is?</p> <p>S: I know that if my dad has brown eyes and he has the dominant chromosome then that covers up my moms recessive chromosomes S: So I would have brown eyes</p> <p>T: something like that, do you understand how that works?</p> <p>S: Well...yeah i just told u</p> <p>T: We will come back to this, what do you understand about mutations?</p>

Table 8.4: Full transcript of an example role-play session (1)

Round 1 chat	Round 3 chat
<p>T: Good morning, so nice to have you in my class today. We are going to be learning about genes and chromosomes today!</p> <p>S: okay</p> <p>T: Do you remember anything about genes and chromosomes from middle school?</p> <p>S: I remember that we have 23 chromosomes that are passed down from our parents</p> <p>T: Great! Do you remember what recessive and dominant alleles are?</p> <p>S: They determine what we look like.</p> <p>T: Great! Is there a difference between the two?</p> <p>S: Something about if one parent has brown eyes which is a dominant color and the other has blue which is recessive then the kid will have brown eyes because thats the dominant color</p> <p>T: Great! You seemed to have really retained that information from middle school!</p> <p>T: Do you remember how those genes are passed down from the parents to the child?</p> <p>S: thank you</p> <p>S: genes are a part of chromosomes</p> <p>T: Okay. How is that so?</p>	<p>T: Good morning! Today we are learning about genes and chromosomes. Do you remember anything about that from middle school?</p> <p>S: we have 46, 23 from each parent</p> <p>T: Okay. Im so happy to hear that you seem to remember things from middle school.</p> <p>T: What do chromosomes do?</p> <p>S: they determine what we look like</p> <p>T: Okay. What do you know about dominant alleles?</p> <p>S: that its one of two options. it beats out the other one.</p> <p>T: Okay. Interesting way of phrasing it. Are there any exceptions?</p> <p>S: something like if the kid has green eyes it a mutation</p> <p>T: Well, I was talking about something a little different.</p> <p>S: well...</p> <p>T: What about recessive genes? Is ever it possible to have a recessive trait? How?</p> <p>S: no dominant beats recessive</p> <p>T: Okay. Interesting point of view. We will learn more about this later and test your theory.</p> <p>S: okay</p> <p>T: How are these genes passed down?</p> <p>S: no clue</p>

Table 8.5: Full transcript of an example role-play session (2)

Chapter 9

Leveraging Community-Generated Videos and Command Logs to Classify and Recommend Software Workflows

Users of complex software applications often rely on inefficient or suboptimal workflows because they are not aware that better methods exist. In this paper, we develop and validate a hierarchical approach combining topic modeling and frequent pattern mining to classify the workflows offered by an application, based on a corpus of community-generated videos and command logs. We then propose and evaluate a design space of four different workflow recommender algorithms, which can be used to recommend new workflows and their associated videos to software users. An expert validation of the task classification approach found that 82% of the time, experts agreed with the classifications. We also evaluate our workflow recommender algorithms, demonstrating their potential and suggesting avenues for future work.

9.1 Introduction

Modern complex software applications often include hundreds or thousands of commands, which further form a much larger number of workflows a user can use. The large variety of commands and workflows raises two issues. First, predesigned tutorials cannot exhaustively cover all the different workflows. Second, users may get stuck in inefficient or suboptimal ways of completing tasks if they are not aware that better workflows exist.

Prior work on software learning addressed this awareness problem [46] at a command-level granularity by recommending individual commands [86, 93], or videos based on command usage [94]. However, command-based recommendations may not consider the user's higher-level workflow needs, or help the user to understand how to use already-known commands in new ways.

Knowing the tasks that a user is working on, and the work-flows they are using, is the first step towards providing better-personalized software learning support. For example, upon recognizing a user's workflow, the application could recommend alternative or more efficient workflows, display sample tutorial videos to help with their task, or provide links to relevant community-created content. By investigating software learning recommendation systems at a work-flow level, this

work complements the existing body of software learning research that focuses on individual commands.

With the above as motivation, this paper contributes a hierarchical approach to mining user workflows at both task and command-set levels. In the first layer of our hierarchical approach, we use Bi-term Topic Modeling (BTM) [150] to infer 18 high-level user tasks (e.g., “Rendering”, “Beginner Sketching”, and “Advanced Surface Modeling”) from command logs associated with 11,713 videos of people using the software. A study found an 82% expert agreement with the algorithm’s classification of videos into task categories. In the second layer, we apply a frequent itemset mining and ranking approach [30] to acquire frequent patterns of commands under each task. For example, a pattern under the task “Beginner Sketching” may look like Center Rectangle, Create Sketch, Sketch Dimension, Edit Sketch Dimension.

Based on this hierarchical understanding of user workflows from command logs, we propose and evaluate a design space of four algorithms which recommend community-generated videos to the user, demonstrating relevant workflows. We evaluate the performance of the four algorithms along the dimensions of relevance and novelty. Users had high ratings on the relevance of the videos, with pattern-based recommendations being more relevant and familiar to the user than task-based recommendations.

Our work contributes a new method to infer user workflows and shows how this method can be utilized to recommend learning videos for a 3D design application. We conclude by discussing how our approach to workflow identification can generalize to other applications and can inform future designs of support systems for software learning.

9.2 Related Work

Our work directly builds upon prior work on tools to support software learning, especially recommender systems and community-enhanced software learning systems. We also draw upon methods used in the literature to mine user data. In this section, we review related work in each of these areas.

9.2.1 Relevant Applications for Software Learning

There have been many efforts in the HCI community to design tools to support the learning and use of software, ranging from reflective visualization tools [14, 85, 91], to tools that provide more active support, such as in-software recommendations and feedback [86, 93, 94]. A range of different channels for providing support have been investigated, from in-application contextual help [45], to community support [83], to auto-generation of demonstration videos [77].

Tools to Raise User Awareness

Visualization tools are designed to raise users’ awareness of their usage patterns and performance, with the goal of encouraging users to adopt more efficient methods. Such visualization tools have been shown to be successful in raising awareness [14, 85, 91], but can be limited in that they reflect existing behavior, rather than providing users with insights into new ways of working. Other

work has adopted more proactive approaches, e.g., CommunityCommands [93] uses collaborative filtering to recommend new commands, and Ambient Help [94] continuously recommends video resources based on the user’s recently used commands.

The above systems operate at the command level. In this work, we also take an proactive recommendation approach, but we develop tools to understand user tasks and recommend personalized resources at a workflow level. Our approach is informed by the collaborative filtering algorithms developed in CommunityCommands [93], and explores a design space for workflow-based recommender systems.

Community Enhanced Learning

Prior work on software learning and “learner-sourcing” [61] has demonstrated the benefits of community-created learning resources, and the potential of repurposing community-created content for software learning purposes. CADament [83] allows players to acquire new skills by observing the workflows of their opponents in a multiplayer game. CoScripter [80] enables end-users to create and share scripts to automate web-based processes. FollowUs [77] enables community-enhanced tutorials, which improve as more users work with them. Techniques have also been explored to extract command demonstrations from workflow videos [76], and to elicit workflow metadata for how-to videos [64].

Motivated by the work above, our approach leverages community-generated workflow videos to model common workflows in an application, and repurposes these videos as a means for presenting recommended workflows to the user.

9.2.2 User Data Mining

Dev and Liu [30] provides a good summary of prior work on user behavior modeling [2, 101, 105], event sequence [107] and clickstream data modeling [133]. In their work [30], they use a frequent pattern mining approach to identify user tasks in a photo editing application from command logs. They also developed a ranking algorithm to select more coherent patterns. Our work directly builds upon this approach, extending it to be more applicable for software with more diverse usage domains, and utilizing it to provide workflow recommendations.

Outside of the software learning literature, topic modeling is a common generative model to extract topics from a corpus [43]. Prior work has successfully applied topic modeling to mine user behaviors from sequence data. Huynh et al. [55] used this approach to identify routine behaviors from sensor data. In this case, a topic is a behavior (e.g., having lunch), and the words are activities associated with this topic (e.g., walking freely, picking up cafeteria food, queuing in the line, etc.) Wen and Rose [144] used a similar approach to identify click patterns in data from Massive Open Online Courses.

In our work, we develop a hierarchical approach combining topic modeling and frequent pattern mining approaches to mine software workflows at two levels: a task level, and a finer-grained command-pattern level.

9.3 Dataset

The software application we target with our approach is a 3D modeling application designed for consumer, commercial and educational use. The application has over 1,000 commands, separated into a set of high-level workspaces (Model, Sketch, Assemble, etc.) Collectively, this rich feature set enables a wide variety of workflows, including modeling, mesh editing, simulation, and animation. Even for a single task, users can take many different approaches. For example, for basic modeling, users can start from primitive shapes (e.g., a box or cylinder), or can sketch in 2D then transform the sketch to 3D using extrude or other operations.

We collected detailed natural usage logs for 20,000 users of the software from June 25 to August 25, 2017, including 255,643 user sessions and 20 million command invocations. In addition to these usage logs, we collected video data from an online community repository where users up-load videos of their usage of the software. These videos have an associated meta-data file with time-stamped command usage data. We collected data for 11,713 videos, with 470,811 commands invoked across these videos. We preprocessed the command logs of the videos to be in the same format as the natural logs from the product, and also collected video attribute data including video length and view count.

9.4 Understanding User Tasks from Videos

The first step to recommend personalized resources to users is to understand what the users are doing in the software. To this end, we developed a hierarchical approach to mine user workflows at both task and command-set levels. In the first layer of our hierarchical approach, we used topic modeling and inferred 18 meaningful user tasks (topics). In the second layer, we mined frequent command patterns for videos of each task respectively, resulting in 233 patterns in total. The rationale for this two-level approach is that simply mining frequent command patterns from the entire corpus of command logs can lead to an over-representation of command patterns for frequently-performed tasks, with those for less-frequent tasks drowned out by the volume of log data for more-frequent tasks. Adding an initial stage of topic modeling allows less frequent but distinct activities to be captured as well.

In this section, we describe the limitations of the state-of-the-art approach [30] to mining frequent tasks on our dataset, and then introduce the first layer of our hierarchical approach – using topic modeling to mine user tasks.

9.4.1 Frequent Pattern Mining Approach

The current practice of identifying frequent user tasks involves applying frequent pattern mining techniques, such as frequent itemset mining and sequential pattern mining [3, 4, 28, 92]. However, such existing techniques do not account for the unique characteristics of software log data. For example, in software logs it is common for users to perform a task by executing a set of operations contiguously with no, or few, outliers. Users may also execute a required operation multiple times within the duration of a task. Recent work by Dev and Liu [30] developed an outlier-based

ranking algorithm to rank frequent patterns mined from user log data, which addressed the above challenges specific to software logs. We adopted their approach as a starting point.

Initially, we applied Dev and Liu’s approach to our dataset without modifications, but found that it mainly identified patterns related to the software’s most frequently-used workspace (Sketching). This suggests that for complex software applications with diverse usage domains, a frequent item-set mining approach may not be sufficient to identify patterns representative of the full range of work-flows in the software.

9.4.2 Topic Modeling Approach

A topic model is a type of statistical model for discovering abstract “topics” that occur in a collection of documents. Although it was originally developed as a text-mining technique, topic modeling has also been used to mine human behaviors [55, 144]. We consider the way a software user composes a session to be similar to the generative process of a document in topic modeling. Users first decide which task to work on, and then choose commands for that task. Additionally, for complex software applications, it is often the case that users will use slightly different command sets to accomplish similar tasks. The relationship between commands may not be captured by frequent pattern mining approaches, since they measure the co-occurrence of a set of commands. We see an opportunity for topic models to capture these relationships.

We define the problem of inferring user tasks from in-situ command logs as an unsupervised machine learning task, due to the lack of ground truth data. In this work, we collected a large dataset of community-generated videos showing various uses of the software, with corresponding command logs. This enables us to first infer user tasks through unsupervised machine learning and then validate the results using the additional context provided by the videos.

We used the command logs from the video dataset as training data for the topic model, treating each video as a document, and each command as a word. After labeling and validating the inferred topics, we applied the topic model to the community logs to classify user tasks for all users.

Data Preprocessing

We extracted the command logs from the video dataset, and filtered out 34 “stop word” commands identified by domain experts, such as “Constrained Orbit”, “Free Orbit”, “Pan”, and “Cancel”. We only included videos with at least 2 unique commands. This resulted in 11,713 videos with 952 unique commands, and 470,811 total command invocations.

LDA vs. BTM

We initially applied a common topic modeling algorithm, Latent Dirichlet Allocation (LDA) [16] (using Gensim [39]) to infer topics from the command logs of the video dataset. To select the number of topics, K , we tested values from 5 to 100 in increments of 5. A researcher and a domain expert analyzed the output for each value of K to identify the most sensible results. New topics (“Animation”) emerged at $K=20$, as compared to $K=15$, and for $K \geq 25$, we started to see overlapping topics, especially related to sketching. We did not further refine the final number of

topics by testing values of K between 20 and 25, due to limitations in time with the domain expert. Based on the above, we finalized at K=20.

A limitation of the LDA algorithm is the sparsity issue. Some videos contain a small number of commands, resulting in a sparse document-word (video-command) matrix. To address this, we adopted the Bi-term Topic Modeling (BTM) approach [150], which is designed to infer topics from short texts. BTM explicitly models word co-occurrence patterns to enhance the topic learning, and uses the aggregated patterns in the whole corpus when learning topics, to solve the problem of sparse word co-occurrence patterns. We applied BTM on the command logs of the video dataset and used a similar approach to tune the number of topics parameter, which also generated optimal performance for K = 20 topics. The researcher and a domain expert did a qualitative comparison of the 20 topics generated by LDA and BTM respectively. We concluded that BTM generated more coherent topics, with fewer overlapping topics, and additional topics not identified by LDA. Thus, we used BTM to classify user tasks. We refer to the topics generated by BTM as “tasks”.

Output of BTM

The output of BTM includes (1) a topic-word (task-command) distribution matrix, and (2) a document-topic (video-task) distribution matrix. Using the video-task matrix, we assigned each of the 11,713 videos as belonging to the “task” (i.e., topic) with the highest weight for that video. We also define the following two similarity terms:

Task-Task similarity—each task is represented in a task-command vector by the topic-word (task-command) matrix. We define the task-task similarity as the cosine similarity between the two task-command vectors. We used a similar definition of task-task similarity as used in Labeled LDA [108].

Video-Video similarity—each video is represented in a video-task vector by the document-topic (video-task) matrix. We define the similarity between two videos as the cosine similarity between the two video-task vectors. Cosine similarity can compare documents in terms of their subject matter [121], rather than length or other attributes. This makes it appropriate for calculating similarity of videos and tasks, which may vary in length but concern the same high-level content. While cosine similarity comes from a different modeling approach than probabilistic modeling, we selected it over probability-based metrics (e.g., Kullback-Leibler Divergence) because it is well-known, easy to implement, and has been found to be as effective as probability-based metrics in applications similar to our own (e.g., filtering redundant documents [155], and computing similarity between user-topic vectors generated by LDA [106]).

Expert Labeling

While the topic modeling algorithm provides an association of commands to topics, it does not provide a semantically meaningful label for these “tasks”. To generate such labels, we recruited two domain experts from the company that developed the software. For each “task”, we showed the top 10 weighted commands for the task, and the top three videos ranked by weight for that task. We included three multiple-choice questions and one open-ended question in the survey to understand whether experts found the “tasks” to be meaningful, and to label each with a name.

Expert Labeling Survey Results

In the open-ended question, we asked the experts to give a name to each task (i.e., what they think the user is working on when seeing these commands used together). The experts' responses are shown in 9.1. Following the survey, a researcher met each expert to discuss their understanding and finalize a name for each task (also shown in 9.1). We dropped one task that was indicated as not meaningful by the experts, and combined two tasks on beginner sketching, resulting in 18 distinct tasks. The number of videos that are categorized into each task is shown in column "Videos".

To further assess the effectiveness of the algorithm, we asked experts to indicate whether the commands composing each task are frequently used together. In a multiple-choice question, we asked each expert to rate "How frequent/likely do you think these commands would be used together?" The experts respectively rated 15/20 and 17/20 tasks as meaningful (9.2). This provides initial validation of our approach of task recognition using topic modeling.

The experts also rated the helpfulness of the videos in deciding the name of the task. Experts found the videos to be helpful or neutral for 16/20 and 16/20 of the tasks respectively. The two experts also showed high agreement on this question, with a Pearson coefficient of 0.48 ($p=0.03$) between their ratings. This is a promising result, indicating that the community corpus of videos can be used to demonstrate new workflows to users as part of a recommender system.

Id	Expert 1	Expert 2	Final Task Name	Videos
1	Intermediate Sketching	Offsetting sketch geometry and trimming lines back that are overlapping	Intermediate sketching (offsetting sketch geometry and trimming lines)	239
2	Advanced Surfacing	Surfacing and Sculpting	Advanced surfacing (surfacing and sculpting)	296
3	Beginners Design	Basic Part Modeling	Basic part modeling	673
4	Advanced sketching	Editing Splines in a sketch	Editing splines in a sketch	293
5	Design (sketch and features)	Extruding text	Creating features from sketches	1251
6	Surfacing	Fixing Surfaces/Patching up surfaces	Fixing or patching surfaces	249
7	Rendering	Add appearances, Rendering a design	Rendering (adding appearances, rendering a design)	166
8	Simulation	Simulation	Simulation	116
9	More aimless clicking	Editing a design	Editing a design	506
10	Drawing Creation	Creating a drawing of a design	Creating a drawing of a design	131
11	Expert Sketching	Creating construction planes and then creating sketches on those planes	Intermediate sketching (creating construction planes)	998
12	Intermediate CAM	Creating CAM tool paths	Intermediate CAM (creating CAM tool paths)	835
13	Someone is aimlessly clicking around	Assembling components	Copy and pasting components, and assembling them	528
14	Industrial Design from image	Inserting a canvas then using a TSpline body to match the canvas	Industrial design based on a reference image	104
15	Animations	Creating an animation	Animation	53
16	Sketching	Constraining a sketch	Beginner sketching (constraining/dimensioning a sketch, fully defining a sketch)	458
17	Sculpt	Editing a TSpline body	Sculpting (editing a T-spline body)	793
18	<i>3rd Party Add In</i>	<i>Not sure</i>	<i>Dropped this topic</i>	<i>445</i>
19	Parametric Design	Dimensioning a sketch	Beginner sketching (constraining/dimensioning a sketch, fully defining a sketch)	1139
20	Assembly	Assembling components	Assembly	583

Figure 9.1: Task names, as labeled by experts, with the number of videos that are categorized for each task.

9.5 Study 1: Task Categorization Validation

The aim of our first study was to evaluate the topic modeling approach for categorizing user tasks. Based on the labels provided by the domain experts, we asked a new set of ten users to watch a selection of videos, and identify the task demonstrated in the video, and the similarity of pairs of videos. The goal was to evaluate whether users’ responses would match the topic modeling algorithm’s categorization.

9.5.1 Video Study Design

Based on several rounds of piloting, we developed the following two question types answered by users in this study.

Labeling Questions

The first question type asked participants to view a single video, and choose a task category for that video from a list of four possible options (9.3). The list of choices for each video was constructed so that it contained the label provided by the topic modeling algorithm, as well as three other tasks. The three additional tasks were selected by ranking the remaining 17 tasks produced by the topic modeling approach by similarity to the task of the target video (using the task-task similarity metric as defined earlier). This ranked list of tasks was divided into three

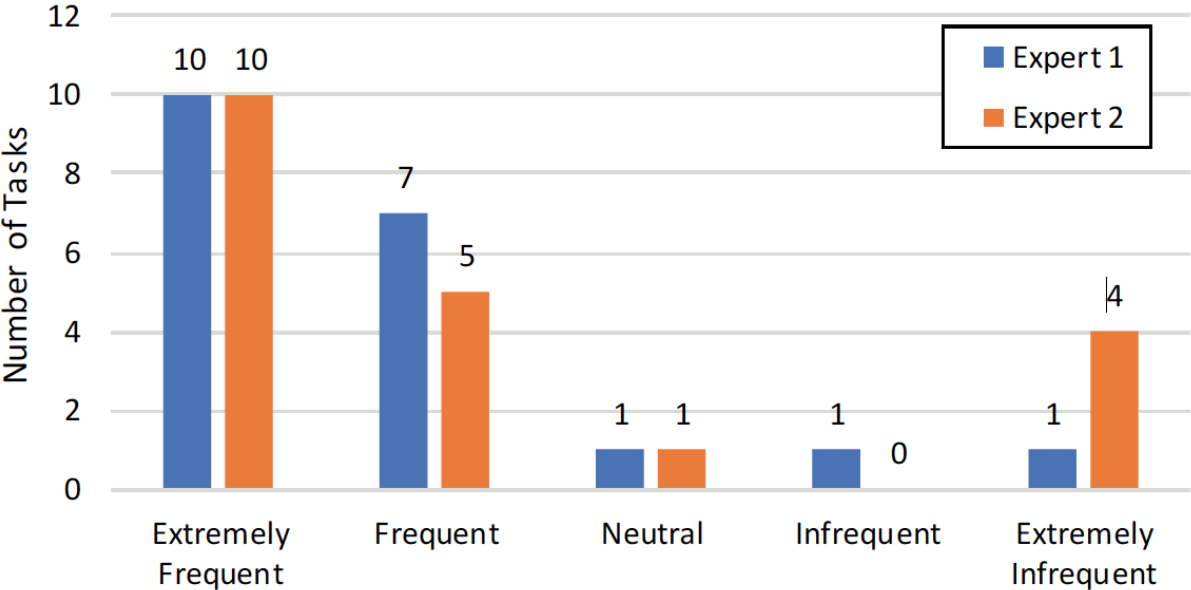


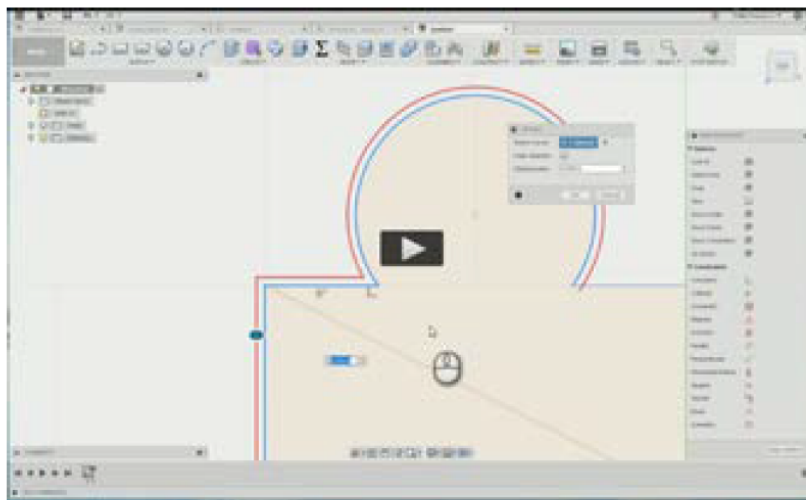
Figure 9.2: Expert ratings for Q1: “How frequently/likely do you think these commands would be used together”

roughly equal-sized tiers, and one task was selected from each tier. This results in each question containing a mix of tasks that our algorithm believes are close to the video, and that are far from it.

For example, the BTM algorithm categorizes the video shown in 9.3 as belonging to “Intermediate sketching (offsetting sketch geometry and trimming lines)”. The three additional choices, ranked by their similarity to this task, are “Beginner sketching” (0.8), “Editing splines in a sketch” (0.25), and “Simulation” (0.12). Using this approach, we include choices that are similar to the task chosen by the algorithm, and ones that are different, without overwhelming the participant with all 18 possible choices. This enables us to investigate whether the distribution of answers over choices is consistent with the similarity between tasks.

Similarity Questions

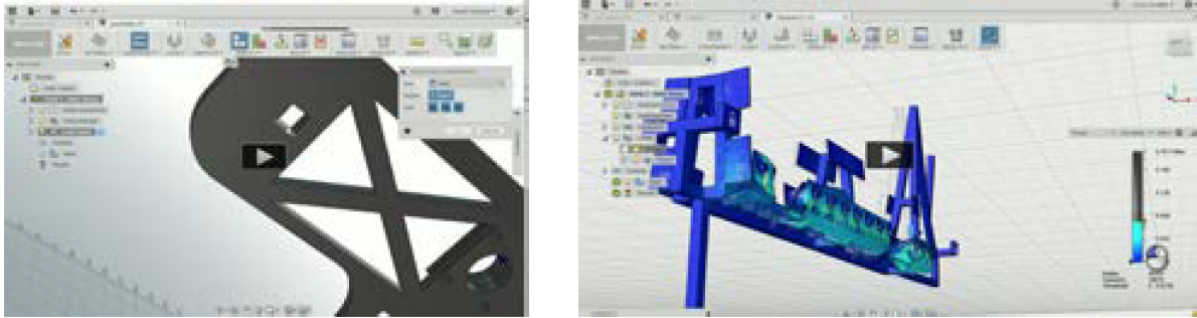
The second question type asked the participant to watch a pair of videos, and evaluate whether they believed the tasks being performed in the two videos were similar or not, on a 5-point Likert scale (9.4). To construct the similarity questions, we randomly selected a video from the dataset as the target video, and ranked all the other videos based on their similarity to the target video (using the video-video similarity metric as defined earlier).



Which of the following categories best describes the task being performed in the video?

- Simulation
- Intermediate sketching (offsetting sketch geometry and trimming lines)
- Editing splines in a sketch
- Beginner sketching (constraining/dimensioning sketch, fully defining a sketch)

Figure 9.3: Example labeling question – Select the category that best describes the task being performed in the video.



To what extent do you agree with the following statement:

The tasks being performed in the two videos are similar.

- Strongly Agree
- Agree
- Neither Agree Nor Disagree
- Disagree
- Strongly Disagree

Figure 9.4: Similarity question – Rate the similarity of the tasks performed in the two videos.

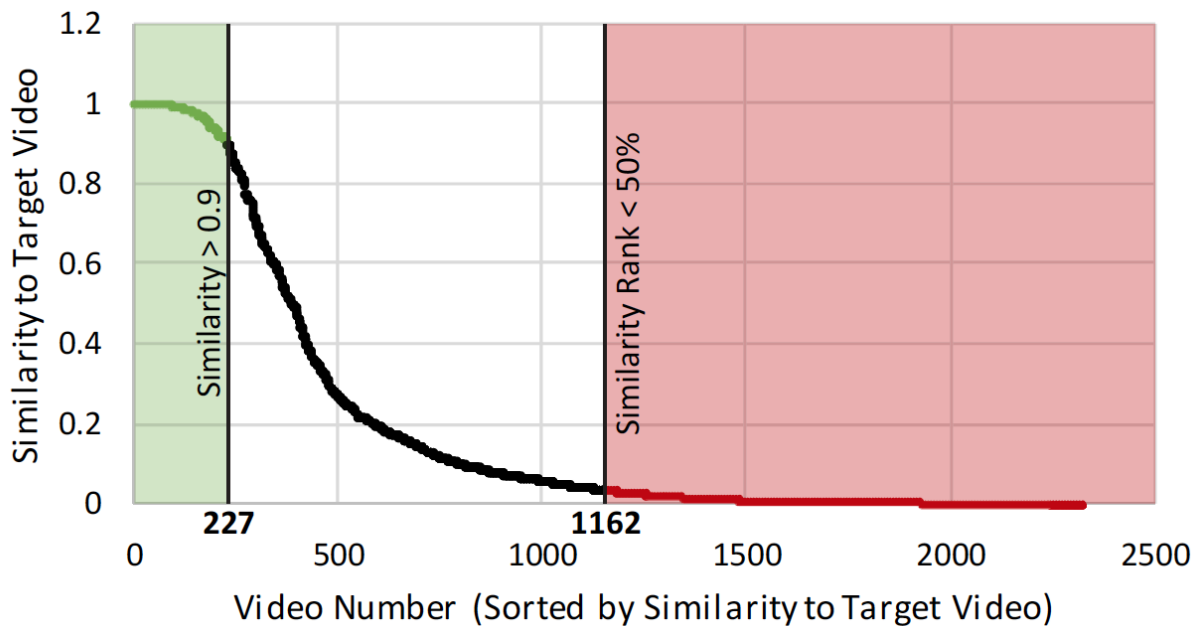


Figure 9.5: Similarity between a video and all other videos. We selected similar videos from the green shaded area, and dissimilar videos from the red shaded area.

The pair of videos shown to the participant was either similar or dissimilar. To form a similar pair, we identified videos with a similarity score higher than 0.9 to the target video (the green shaded area in 9.5), and randomly selected one such video. To form a dissimilar pair, we randomly selected a video from the bottom half of all videos, ranked by similarity to the target video (the red shaded area in 9.5). Our approach for selecting similar/dissimilar items was developed in an ad-hoc manner, based on experimentation with our dataset. In particular, we found that a simpler criterion (e.g., selecting the top 10% of videos, ranked by similarity [93]) did not account for tasks with few videos. For example, the Animation task included only 53 videos, so selecting 10% of all videos ranked by similarity to a video on Animation would select many videos outside this topic. Using a threshold on similarity score avoided this problem.

Participants

We recruited 10 users (8 male, 2 female) that self-identified as intermediate or expert-level users of the software. Participants were given a \$25 gift card for participating.

Study Design

Each participant answered 11 labeling questions, and 18 similarity questions (including 9 video pairs that our algorithm indicated were similar, and 9 that our algorithm indicated were dissimilar), covering a total of 47 videos across 29 questions. In order to cover a larger variety of videos, we designed the study so that 2 participants worked on questions with the same set of 47 videos, with 5 sets of 47 videos in total. We also made sure videos of each task were equally represented in the questions. In total, we covered 235 videos in the study, which ranged in duration from 10-60 seconds. To counterbalance, in each set, we reversed the order of similarity and labeling questions for the two participants. Questions within each category were presented in a random order. To make sure the participants watched the videos and treated the questions seriously, we asked participants to provide a short text justification of their response to each question.

9.5.2 Quantitative Results and Analysis

For similarity questions, if the participant answers “Strongly Agree” or “Agree” for a similar pair of videos, or “Strongly Disagree” or “Disagree” for a dissimilar pair of videos, we count the answer as consistent with the algorithm, otherwise as inconsistent. Participants’ overall agreement on similarity questions with the algorithm was 71%. We also computed the users’ average rating for similar pairs (1.4) and dissimilar pairs (3.4), where the rating is computed on a 1-5 scale, with 5 meaning “similar” and 1 meaning “dissimilar”. For the labeling questions, participants agreed with the algorithm’s classification 82% of the time. The performance of each participant is shown in 9.6.

From the results, we found that participants had a high level of agreement with the algorithm. The main source of inconsistency was the lack of agreement when our algorithm considered two videos to be similar – participants were more conservative about judging two videos to be similar.

In particular, participants had high agreement with the algorithm on labeling questions. As shown in 9.7, 82% of responses matched the algorithm’s classification. Moreover, approximately

half of the responses that did not match the algorithm were in Tier 1, which is the closest to the classification of the algorithm without being an exact match. Overall, 91% of responses were either exact matches or in Tier 1.

9.5.3 Qualitative Analysis of User Feedback

To better understand the circumstances under which participants agreed or disagreed with the algorithm’s classification, we examined the justifications they provided for their ratings. In general, we found that participants were able to give detailed descriptions of the tasks being performed in the videos. Some examples of justifications are provided below:

“They are both linked with motion, both using joints to drive the parts or restrict movement.”

Video Set	Similarity Agreement	Rating for dis-similar pairs	Rating for similar pairs	Labelling Agreement
1	83%	1.3	4.2	82%
	78%	1	3.4	82%
2	61%	1.3	2.4	91%
	72%	1.4	3.1	73%
3	78%	1.1	4	100%
	94%	1.2	4.3	91%
4	61%	1.7	2.8	64%
	67%	1	2.6	73%
5	61%	2	3.6	91%
	56%	2	2.8	73%
TTL	71%	1.4	3.3	82%

Figure 9.6: Summary of results for Study 1, grouped by video set.

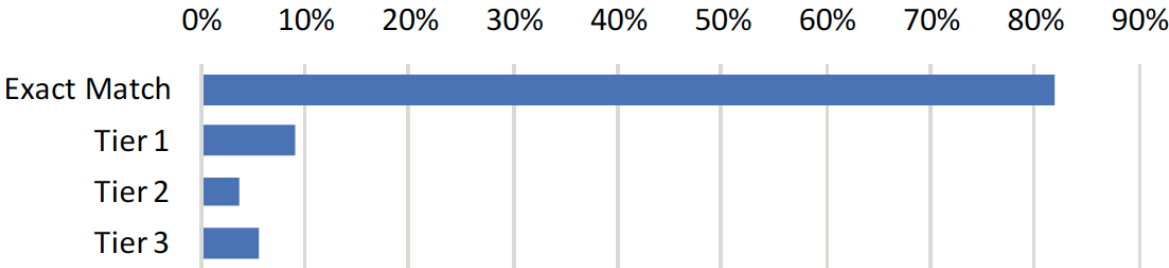


Figure 9.7: Summary of agreement between the algorithm and participants’ responses to the labeling questions.

“Both are short videos that go into the render environment and start in-canvas rendering.”

“Fairly advanced sketching manipulations, all 2D. 2nd video, manipulating 3D assembly with joints and alignments, no actual changes to the geometry just their orientation.”

“The first video is using a sketch on a plane to use as a cutting tool to split a body - the second uses the time line and preferences to modify an existing model.”

Participants’ justifications indicated a range of different standards for judging similarity. While we asked participants to judge based on whether the two workflows were working toward a similar goal, even if their individual approaches were different, or the end results were different (e.g., one succeeded while the other failed), we observed that many participants judged similarity based on other standards, such as the expertise level of the approaches shown in the video, or the specific operations used (e.g., “Both navigate the design space and add features. The former looks unprofessional, but latter looks very skilled.”, and “Both videos create a sketch, but other operations are different”.) Participants also mentioned that their answer could go either way, depending on how similarity was defined (e.g., “It depends on how vague you want to go with the similarities – you could say that they are using some type of constraint by aligning faces or changing dimension etc. – but it is vague”). Further, some participants were particularly strict when judging similarity. In the examples below, participants’ justifications for their ratings suggest an agreement that the tasks had a similar goal, but their ratings do not reflect this:

“1 is creating a drawing view. 2 is editing a drawing view’s scale. The general end goal is to end up with a drawing.” (Rated as “Neutral”)

“Both attempt to simulate how parts work in the physical world” (Rated as “Extremely different”)

“1 is creating a body from a TSpline. 2 is creating variations of TSpline bodies and ends up with bodies.” (Rated as “Neutral”)

Overall, our study results suggest that the BTM algorithm does infer meaningful user tasks that are consistent with the understanding of experienced users of the software. This is encouraging evidence for the value of topic modeling approaches for software log data. Our findings also indicate the value of gathering a corpus of video and associated log data for a software application, as it can be used to validate approaches for modeling user tasks from log data.

9.6 Hierarchical Task Identification

Study 1 validates the first layer of our hierarchical approach, with which we can infer the high-level tasks (topics) the user is working on. However, even when two videos are of the same high-level task, they may contain very different command sets. The second layer of our approach allows us to acquire finer-grained command sets under each task.

We began with the output of BTM, which assigned each video to a task (i.e., the task with the highest weight for that video in the video-task matrix). For the set of videos under each task, we applied the FP-Growth algorithm (using SPMF library [84]) on the command logs to identify frequent patterns. We set different thresholds for each task based on how many videos there were – our rule of thumb was that the number of frequent patterns acquired for each task should be within the range of 5-10% of the total number of videos for that task. For the patterns acquired under each task, we applied the ranking algorithm developed by Dev and Liu [30] and set the minimal

length for a pattern to be 3 and the cutout cohesion score to be 2. By choosing cohesion score of 2, we allowed 1 outlier for a pattern with 3 commands. For example, the pattern Construct Sketch, Draw Line, Add Geometry Constraint to Sketch contained three commands and had a cohesion score of 2, because for the 206 times that this pattern appeared, at least half of the times there was another command that appeared in the sequence other than the three commands in the pattern (i.e., an outlier). Examining the video command logs, there were cases where the three commands appeared together with no outliers, and other cases such as Draw Line, Construct Sketch, Trim Sketch, Add Geometry Constraint, and Draw Line, Construct Sketch, Add Tangent Handle, Add Tangent Handle, Add Geometry Constraint where this was not the case. Setting the cutout cohesion score to 3 would result in a loss of such length-3 patterns that appeared frequently with 1 outlier in between the commands. We refer the reader to Dev and Liu [30] for additional context surrounding our choice of cohesion score and allowing outliers. Using this approach, we got 233 frequent patterns in total for the 18 tasks. The final distribution of command patterns by task is shown in 9.8.

Comparing this hierarchical approach with simply applying the above command-set identification to all data, we found greater diversity in the tasks identified. Specifically, with-out first

Task Name	Count
Intermediate sketching (offsetting sketch geometry and trimming lines)	9
Advanced surfacing (surfacing and sculpting)	1
Basic part modeling	0
Editing splines in a sketch	19
Creating features from sketches	49
Fixing or patching surfaces	4
Rendering (adding appearances, rendering a design)	6
Simulation	34
Editing a design	9
Creating a drawing of a design	12
Intermediate sketching (creating construction planes)	13
Intermediate CAM (creating CAM toolpaths)	10
Copy and pasting components, and assembling them	4
Industrial design based on a reference image	2
Animation	1
Sculpting (editing a T-spline body)	2
Beginner sketching (constraining/dimensioning a sketch, fully defining a sketch)	57
Assembly	1

Figure 9.8: Distribution of patterns by task.

applying BTM, all 53 frequent patterns acquired were for sketch-related tasks.

Examining the output of our approach indicated that it provided reasonable results. For example, for the Beginner Sketching task, we found patterns such as Center Rectangle, Create Sketch, Sketch Dimension, Edit Sketch Dimension showing the user created a sketch, drew a rectangle, and edited its dimensions. For the Intermediate Sketching task, we found patterns such as Line, Extrude, Stop Sketch, Trim showing the user is drawing and trimming lines in a sketch, and extruding from the sketch. For the Assembly task, we found patterns such as Activate Environment, Joint, Drag Joint Origin which shows the user activated the workspace, dragged the joint origin and then made a joint.

9.7 Recommender System

Using the hierarchical approach, we trained a topic model using BTM on the command logs of 11,713 videos, to infer 18 topics representing high-level tasks in the software (first layer), and then acquired frequent patterns for each of these tasks (second layer), resulting in 233 patterns in total. These topics and command patterns allow us to infer a task distribution for each user, and to characterize users and videos based on which of the patterns are exhibited in their log data. Specifically, we used this model of tasks and command patterns to design and implement four collaborative-filtering algorithms to recommend workflows and associated videos.

To form the community for collaborative filtering, we sampled 20,000 users and collected their log data for the period of June 25 to August 25 of 2017. In all, this included 255,643 user sessions and about 20 million command invocations. We applied the model trained using BTM in the first layer of the hierarchical approach on the community data to infer task usage for each user, resulting in a user-task distribution matrix. We then searched for the appearance of each of the 233 patterns acquired from the second layer of the hierarchical approach in the command logs of each user session. This results in a user-pattern frequency matrix.

The recommender system works in three steps. First, for a given user, the system gets his/her task and command pattern usage from the above matrices. Second, it selects either a task (first layer) or a command pattern (second layer) to recommend using collaborative filtering. Finally, it selects a candidate video to recommend based on the chosen task or pattern.

9.7.1 Recommender System Design Space

Based on this hierarchical understanding of user workflows both at a task level (inferred from the topic modeling approach), and at a pattern level (inferred from the frequent pattern mining approach), we propose and evaluate a design space for workflow-based video recommender systems. The first dimension of our design space is granularity, basing our recommendation on either a topic level or a pattern level. Videos that are recommended at a topic level target a general task, e.g., sketching, whereas videos that are recommended at a pattern level target a specific pattern of commands, e.g., drawing a line, applying a constraint, and editing dimensions. The second dimension we evaluated was topic relevance. Recommendations were either most-familiar topic (MFT), meaning the recommended videos match the user's most commonly-used topic, or less-familiar topics (LFT), meaning the recommended videos are outside of the user's most

	Most-Familiar Topic	Less-Familiar Topics
Topic Level	1. <i>Topic-MFT</i>	2. <i>Topic-LFT</i>
Pattern Level	3. <i>Pattern-MFT</i>	4. <i>Pattern-LFT</i>

Table 9.1: A design space for workflow recommender systems.

commonly-used topic. The intention is for the MFT recommendations to be more familiar and relevant, and for LFT recommendations to be more novel. Exploring this dimension allows user to explore the tradeoff between relevance and novelty [82]. Based on these two dimensions, we proposed a design space of four workflow based video recommendation algorithms (Table 9.1).

9.7.2 Recommendation Algorithms

All four algorithms make recommendations in two stages. First, a topic is selected (topic-level algorithms), or a set of five patterns¹ is selected (pattern-level algorithms). Second, based on the selected topic or patterns, videos are chosen that either belong to the topic, or contain the selected pat-terns. The following sections describe the specific algorithms.

Algorithm 1: Topic-MFT

Step 1: Compute the task distribution for the target user.

Step 2: Choose the task the user has most frequently used. For example, in 9.9 we visualize the task distribution for a target user. In this case, we would select Task 5.

Step 3: Select five videos for the chosen task. We select all videos in the dataset that belong to this task, and compute the similarity between the target user’s task usage with all videos selected. Videos with a similarity greater than 0.9 are ranked by their view counts, and the top five videos

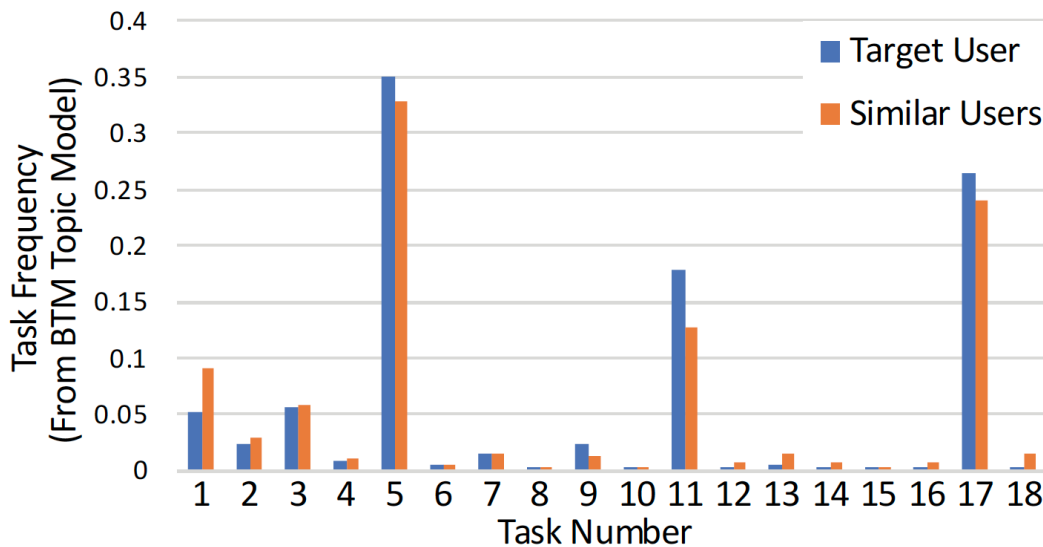


Figure 9.9: Example task distribution, user vs. similar users.

are selected. The threshold 0.9 followed a similar rationale as introduced in Study 1. The task similarity step guarantees that the videos will be close to the user’s typical workflows, and the view count ranking ensures that we select higher-quality videos.

Algorithm 2: Topic-LFT

Step 1: Compute the task distribution for the target user and all 20,000 other users in the community.

Step 2: Find similar users. We select users that have a task similarity score larger than 0.9 with the target user. Task similarity here is defined as the cosine similarity between the user-task vectors. 9.9 shows the comparison of task distributions between a target user and similar users.

Step 3: Compare target user to similar users. We compute the task weight difference between the target user and similar users, and select the task that has the largest delta between similar users and the target user. For the example in 9.9, Task 1 would be recommended. We then use a similar algorithm as in Approach 1, Step 3 to select five videos.

Algorithms 3 & 4: Pattern-MFT, Pattern-LFT

Step 1: Compute the pattern frequency distribution for the target user, and all 20,000 other users in the community.

Step 2: Find similar users based on pattern frequency. Similarity here is defined as the cosine similarity between user-pattern frequency vectors. Since the pattern frequency similarity is much lower than task similarity, it was difficult to determine a threshold for selecting similar users. We chose N=200 to select the top 200 users based on the ranking of pattern frequency similarity with the target user.

Step 4: Compute expected pattern frequency for the target user. To calculate the expected frequency for each pattern, we follow the method used by Matejka et al. [93]. We define the expected frequency, ef_{ij} , for pattern p_i and user u_j :

$$ef_{ij} = \sum_{k=1}^n w_{jk} pf_{ik}$$

, where w_{jk} is the similarity between u_j and u_k , and pf_{ik} is the frequency of pattern p_i for u_k .

Step 5: Remove previously used patterns. We then rank the patterns based on the expected frequency and remove patterns that the target user has been observed using.

Step 6 (Algorithm 3): Select patterns for the user’s most relevant task. In the list of patterns ranked by expected frequency, to select patterns that are more relevant to the user, we select the top five patterns that belong to the most frequently used tasks by the target user.

Step 6 (Algorithm 4): Select patterns of less relevant task. In the list of patterns ranked by expected frequency, to select patterns that are less relevant to the user, we select top five patterns that are outside the user’s most frequently used tasks (as defined in Algorithm 3).

Step 7: Choose videos based on patterns. For each chosen pattern, we first select all videos that contain that pattern. We then rank the selected videos based on their task similarity to the target user, and rank the top 10 by view count. Finally, we select the top-viewed video as the video

for that pattern. This is the same method for video selection used in Approach 1 and 2, which is designed to guarantee the video is close to the user’s typical workflows and of reasonable quality.

9.8 Study 2: Workflow Recommendations

To evaluate the proposed algorithms, we conducted a study where we generated a personalized set of videos for participants, and sent them a survey where they could view the videos and rate the recommendations. This follows the methodology used in past work by Matejka et al. [93].

9.8.1 Participants and Procedure

We recruited 8 users that had actively used the software during the past 2 months (1 female, 7 male). With permission, we retrieved participants’ natural log data for the past two months, from June 25 to August 25, 2017. We made 20 video recommendations in total for each user – five videos from each of the four algorithms described above. We filtered videos to be shorter than 5 minutes. The videos were presented to participants in random order.

For each video, participants were asked to rate to what extent they would agree with the following statements (1=Strongly Disagree, 5=Strongly Agree): (1) I was familiar with the workflow (or workflows) shown in this video. (2) I may use the workflow (or workflows) shown in this video. (3) This video would be a good demonstration for someone who was unfamiliar with the workflow (or work-flows) being shown. The first question (Familiarity) was used to evaluate whether the workflows were novel to the user. The second question (Relevance) was used to evaluate whether the workflows were relevant to the user. The third question was used to evaluate the quality of the video, and the feasibility of using community generated videos for learning new workflows. Each question was followed by a justification text box.

9.8.2 Quantitative Results and Analysis

9.2 shows the average rating of each recommendation algorithm on the two dimensions of relevance and familiarity. In general, our results indicate that participants found the recommended videos to be relevant (indicating that they would use the workflow), and familiar (indicating lower novelty). Given the low sample size ($n=8$) it is difficult to make definitive conclusions, however we do see potential trends of higher familiarity ratings for pattern-level recommendations than topic-level recommendations ($p=0.08$) and higher relevance ratings for the pattern-level recommendations as well (ns). Our interpretation of pattern-based algorithms generating more familiar and relevant recommendations is because similar users, as identified by pattern frequency similarity, are more likely to use similar patterns. Even if previously-used patterns are removed, the patterns recommended may be of a similar general task.

While the LFT approaches showed some potential impact on the novelty of the recommendations, the novelty ratings were lower than we expected. This could be because we favored relevance in the design of the recommendation algorithms. For instance, in the video selection step, we select-ed videos based on the similarity between the target user’s task usage and the task

Algorithm		Relevance Rating	Familiarity Rating
Topic-Level	Most-Familiar Topic	4.13	3.70
	Less-Familiar Topics	4.10	3.58
Pattern-Level	Most-Familiar Topic	4.23	4.08
	Less-Familiar Topics	4.18	3.95

Table 9.2: Study 2 results: users’ rating on the relevance and familiarity of the recommended tutorial videos.

distribution of the videos, which can cause the recommended videos to be closer to the user’s typical workflows. We revisit this issue in our discussion of future work.

In terms of the video quality, the ratings were generally positive, with an average rating of 3.5 for all the recommendations. In their free-form feedback, participants showed a strong preference for videos with audio.

9.8.3 Qualitative Analysis of User Feedback

Through a qualitative analysis of user feedback, we found that users may rate the videos as very familiar, even if they disclosed in their justification responses that they were only partially familiar with the workflow. Part of this issue is that each video may show more than one workflow or command pattern. If part of the video shows a general task that the user is familiar with, the user may rate the video as familiar, even if a subsequence of the video was novel.

Despite the low level of reported novelty, users did report positive attitudes towards the recommended videos and expressed that they would want to try out the workflows in the future. For example, P1 stated: *“I knew how to use all these features, but hadn’t really thought of using them in combination this way before. The workflow will be useful in the future.”*

P2 expressed that he/she is partially familiar with the work-flow recommended: *“I am partially familiar with the patch workflow but understand how to use extrude to make a groove to a box. I want to learn more about patch and the video gave a good description of how it can be used.”*

P3 was excited about one of the recommended workflows: *“Amazing workflow, I was never familiar with such an approach. Maybe because I’ve never used Remake. I would love to give it a go. This has a lot of uses and potential in the field I’m working, so I would definitely use it.”*

In conclusion, the algorithms show a strong potential for recommending learning resources that are relevant to the goals of the target user. The study reinforces the tradeoff between the two factors of relevance and novelty in our design space. We see opportunities for improving the algorithms and adjusting the design decisions to make more novel recommendations, which we will discuss as future work.

9.9 Discussion and Future Work

A diagram summarizing our approach is shown in 9.10. The overall idea is to first use topic modeling to segment logs into mutually exclusive sets based on high-level tasks (Layer 1), and then to apply frequent pattern mining to each set, to identify finer-grained patterns of command

usage (Layer 2). The resulting topics (i.e., tasks) and patterns are then used as input to software support systems, e.g., recommender algorithms, as we have demonstrated.

While we found that Bi-term Topic Modeling and Dev and Liu’s algorithm for frequent pattern mining were effective, we see the hierarchical approach as being largely independent of these specific techniques, or the assumptions used to tune them to our dataset. In particular, a key finding from this work is that it is valuable to first use topic modeling to segment logs, and then to apply pattern mining to the resulting segments, because it prevents frequent user activities (e.g., sketching activities for our application) from drowning out other distinct activities performed in the software.

The remainder of this section discusses opportunities for future work to develop and build on this approach.

9.9.1 Incorporating Heuristics Based on Expertise Levels

In our approach, we tried to minimize human input. Apart from the expert labeling of topics, the process is data driven. However, we see opportunities to incorporate heuristics to enable more meaningful workflow recommendations (an approach used at the command level in [93]). Though we did not report on it, we had experts evaluate the expertise levels of the 18 tasks produced by our topic modeling, and this data could be incorporated into a workflow recommender system (e.g., to recommend intermediate sketching workflows to users who have been observed doing beginner sketching).

9.9.2 User In-the-Loop Recommender Systems

In this work, we did not apply filtering to control the quality of community-generated content, but we see the potential of integrating such quality-control methods (e.g., machine learning methods to predict video quality [76]). Prior work has shown that “learner-sourcing” systems can harness input from learners to improve the quality of content over time (e.g., ask learners to label activities performed in MOOC videos [62]). Similar approaches could be used to refine the recommendations made by a workflow recommender system, so that recommendations improve over time.

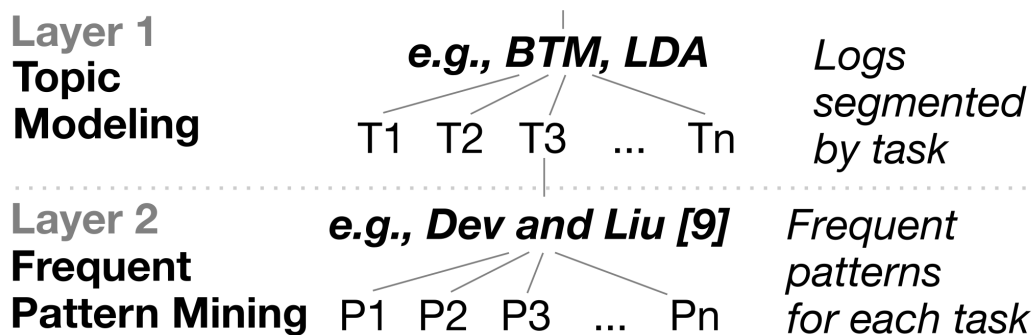


Figure 9.10: Summary of our hierarchical approach.

9.9.3 Generalizability

Though we developed our hierarchical approach for a specific 3D design application, the approach can be applied to other applications as well. Our approach is based on command log data, which is commonly logged in feature-rich software. More unique is that we also leverage data from a user-generated video repository, where the videos are supplemented with command log data. Software companies looking to apply our approach could curate such marked-up video repositories with existing tools, e.g., Autodesk Screencast’s public SDK [12]. Alternatively, prior work has demonstrated approaches to extract command data from existing video repositories [13, 64]. In this way, our approach can be generalized to other applications and software domains.

9.9.4 Limitations

A limitation of our work that could impact its generalizability is that we used heuristic or ad-hoc approaches to choose some parameters (e.g., the cosine similarity threshold of 0.9, and the cutout cohesion score of 2), which would need to be adapted for other data sets. More broadly, the use of cosine similarity is a limitation as it comes from a non-probabilistic modeling approach, and thus it would be valuable to investigate probabilistic-based similarity metrics, such as Kullback-Leibler Divergence [90], in future work. When selecting videos to recommend, we also used an ad-hoc method to select videos that are similar to a user’s typical workflow, which favored “Relevance” over “Novelty” in our design space. Future work could investigate more rigorous approaches to tuning the similarity threshold, or more generally modeling the similarity/novelty of workflow recommendations.

9.10 Conclusion

In this paper, we have proposed a hierarchical approach to classifying user workflows by first applying topic modeling to identify high-level tasks, and then applying frequent pattern mining to identify distinct command patterns for each task. An evaluation showed encouraging evidence that topic modeling can effectively categorize logs into meaningful high-level tasks. As well, the hierarchical approach appears to help identify a larger variety of distinct command patterns. Based on this approach, we proposed a design space of workflow-based recommender systems. An evaluation of four such algorithms was encouraging, and suggests that this approach has the potential to effectively support software users.

Chapter 10

List of Contributions

I summarize the contributions and insights offered by this dissertation below. This dissertation contributes to the literature of Human-Computer Interaction, Learning Technologies, and Learning Sciences.

10.1 Human-Computer Interaction and Learning Technologies

- **This dissertation contributes a novel technique that uses student solution examples to semi-automatically generate authentic multiple-choice questions for deliberate practice of higher-order thinking in varying contexts.** With two classroom experiments, we show that the deliberate practice opportunities created with this technique help students gain conceptual understanding more efficiently and help improve the quality of student open-ended work.
- **This dissertation contributes insights about developing effective learning at scale systems by leveraging the complementary strengths from peers, experts, and machine intelligence, differentiating it from existing systems that solely rely on machine or crowds. There are three components that contribute to the effectiveness of this learnersourcing technique. First, instructors are not good at creating distractors, and actual elaborated student errors are helpful as sources. Second, we apply simple natural language processing techniques to select distractors of interest. Third, we involve instructors closely in the process to enhance quality.**

Building upon prior work on automatic question generation for educational purposes[75], techniques in prior work 1) target lower cognitive skills, such as fact questions and fill-in-the-gap questions 2) does not provide meaningful feedback 3) are domain-specific when leveraging existing language ontology. The strategy these approaches primarily use is limited in that they tend to simply transform given text from declarative statements to questions. In this work, we explore data inputs that are not declarative statements, but elaborated solutions from students that display common misconceptions with accompanying

thought processes. Some of the thought processes students displayed in their open-ended work are leveraged as “feedback” for the new questions. The nature of the data input provides sources for feedback.

On the other hand, we involve instructors at multiple stages. Instructors are asked to specify question creation schemata that target higher order thinking, e.g., evaluation. The re-organization of student open-ended solutions provides examples of varying contexts. This makes the creation of question schemata that target higher order thinking feasible, e.g., having students analyze and evaluate examples. This would not be possible without examples of varying contexts as the input. Instructors also review questions and provide feedback in the end, which enhances the quality of the questions produced. The nature of the data source and the involvement of experts at multiple stages differentiates the techniques UpGrade from existing work in the literature on automatic question generation techniques. Building upon prior work on learnersourcing and crowdsourcing, prior work leveraged learner data such as video watching traces [63], video annotations [64, 89, 143], or explanations [146] to benefit future learners. I explore written homework assignments as a new and powerful input for learnersourcing. Peer review [73] is another direct ‘scaling’ approach for education. For peer review to work effectively, it requires hard rubric work that instructors can’t or don’t want to do and timely and high quality feedback – while not impossible – is extremely rare [137]. In UpGrade, instructors play an important role in quality control compared to prior crowdsourcing/learnersourcing systems and peer review systems. Instructors’ effective participation makes sure the content created is of high quality.

- **Applying the workflow of UpGrade in practice across courses demonstrates the generalizability and practical value of this approach and helps inform the design of an interface to facilitate the independent use of UpGrade by instructors for authoring practice questions.** Over the past few years, we have applied the workflow of UpGrade to 9 modules on topics of research methods and design including heuristic evaluation, survey design, usability findings from think-aloud studies, affinity diagrams, interview question design, creating storyboards for speed dating studies, log data analysis and visualizations, theoretical cognitive task analysis, and performing error analysis with machine learning predictions. Some modules were repeatedly used in the same class and some are used across classes at CMU. For the development and use of these modules, we followed the workflow of UpGrade using a combination of offline, i.e., meeting between an engineer (myself) and the instructors and online, algorithms of UpGrade, approaches. These modules have different types of data input, and we also used different quality control methods for different modules. The effort required from instructors for these modules is 2 hours at most. In this process, I’m also trying to understand instructors’ preferences and traditional workflows in writing practice questions to inform the design of the interface of QuizMaker, with which instructors can independently use the workflow of UpGrade. On a practical side, this past process demonstrates the generalizability of the workflow of UpGrade, with the facilitation of a learning engineer. On the other hand, we leave it to future work to explore the usability and effectiveness of the QuizMaker interface for independent use.
- **When instructor efforts are not available for reviewing and revising the questions to enhance quality control, we demonstrate an effective quality control method using**

psychometric approaches to automatically select high-quality question items from a large question pool. We show that with this quality control method, UpGrade produces a question bank that exceeds reliability standards for classroom use. We demonstrate that crowd (such as MTurk) can be leveraged as a source for quality control. Crowd generates consistent performance as real students. We also demonstrate that when reducing the sample size, false negatives increase and false positives remain the same. This suggests that with techniques such as UpGrade, which produces a large question pool, the system can use relatively small data sets to prune out unreliable items without worrying about having false positives.

- **I demonstrate two examples where complementary human and machine intelligence are leveraged to create educational materials at scale. In both cases, crowds (e.g., past students) offer a powerful data source with structure examples that show common errors. Machine automatically selects, filters and reorganizes examples. Experts (instructors) check information accuracy and comprehensiveness to enhance content quality.** First, in the example of UpGrade, with instructors alone, it is often hard to write lots of elaborated good examples and wrong answers. With machine alone, we see that existing question generation techniques only produce fact questions and are not flexible. Here, past peers offer a powerful data source with structured examples that show common errors. With this better data input, machine can auto select/filter and reorganize examples. Instructors, in the end, will check information accuracy and comprehensiveness. This altogether supports quality content creation at scale. Second, I present a workflow categorization and recommendation technique where I harness examples from online videos and logs to support the use of complex graphical software. Here, peers who can be users of the software everywhere in the world contribute demonstration videos. We collect these videos (and their command logs) as input and use machine to identify an end user’s workflow; expert input is elicited in this process. With this approach, existing online repositories are repurposed as targeted tutorials for end users.

10.2 Learning Sciences

- **This dissertation suggests another layer to further distinguish knowledge components, by the required generation and evaluation efforts in problem-solving. The practical implication for a more nuanced understanding of knowledge components (KCs) is to help instructors make more nuanced and accurate instructional decisions, e.g., using “evaluation-type” exercises for evaluation-heavy skills.** As defined in the KLI framework [70], students gain knowledge components (KCs) through learning events, which can be inferred from performance on assessment events. The KLI framework suggests that kinds of KCs drive instructional event choices. For example, instructional approaches that emphasize recall and spacing of practice may benefit learning of historical facts, vocabulary; whereas instructional approaches that prompt self-explanation in students would be more valuable for learning complex principles, such as Newton’s laws. In my work, we identify problem-solving skills that require critical evaluation efforts (such as heuristic

evaluation, survey question design, etc.). The practical implication for a more nuanced understanding of knowledge components (KCs) is to help instructors make more nuanced and accurate instructional decisions, e.g., using “evaluation-type” exercises for evaluation-heavy skills.

- **This dissertation indicates that, at least for some domains, 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.** Prior work has investigated the use of “evaluation-type” exercises in other age groups and other domains. For example, Yannier *et al.* 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 [152]. Ericsson *et al.* 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. My work adds to this body of literature, focusing on different domains and contexts. The practical implication for Learning experience (LX) is that LX designers may consider, with less guilt, using 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.
- **This dissertation provides further evidence that instructors have so-called “expert blind spots”, revealed through cases where their beliefs and student performance do not match. Specifically, instructors believe open-ended assignments to be better for student learning, which is contradicted by student performance data.** My work provides further evidence that instructors have so-called “expert blind spots”, revealed through cases where their beliefs and student performance do not match [98, 99]. Instructor beliefs are important because they will influence the design of the curriculum and learning experience of students. In both this and a past case [69], we see experts have good reasons for their beliefs, yet data suggests otherwise and a deeper analysis explains why. **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.**
- **This dissertation also makes suggestions to the model of take-home assignments used in a higher-education context, especially relevant to topics similar to the ones we have investigated. We surface an issue that open-ended work students turn in are of low quality, suggesting there are cases when students are not ready and need scaffolding before they do complex open-ended work. An alternative model would be giving students deliberate practice opportunities before the assignment of flexible open-ended work. The deliberate practice opportunities can be easily created with techniques such as UpGrade.** Open-ended assignments are frequently used in colleges and are treated

as a major source of practice and learning in many courses, considering a regular course that is expected to take a student 12 hours each week. Besides the time to attend lectures and complete readings, students are usually expected to spend at least 6-9 hours each week on their assignments. This is about 50% - 75% of students' learning time. One reason for using open-ended assignments is that students will receive individualized feedback on their learning. However, we found the feedback students were given hasn't been satisfactorily informative. The frequency of the same mistakes among students is high, and the persistence of the same mistakes within a student is also high. We demonstrate that it is valuable to provide more instruction before asking students to spend significant time working on their projects.

- **I present the design of a system that benefits from the distinction between generation and evaluation efforts during problem-solving. ELK (Eliciting Learner Knowledge), is a text-based role-playing system that enables pre-service teachers to practice questioning moves through simulated “teacher-student” conversations.** To facilitate “evaluation-type” exercises, we introduce a “Coding” activity in ELK where players evaluate authentic transcripts generated from past role-play sessions. We show that evaluating authentic transcripts generated by others helps participants develop questioning moves to a similar degree as generating improvisational questions in the role-play chat. The “Coding” activity has practical benefits as it can be performed by a single participant alone, which serves as a viable supplementary activity for online role-play systems.

Chapter 11

Discussions and Future Work

In the first section, I talk about the future work specifically relevant to creating and providing deliberate learning opportunities as done by UpGrade. In the subsequent sections, I discuss my observations working on this dissertation and future work I'd like to explore related to human-AI collaboration, learning technologies, higher education, and creativity.

11.1 Creating and Providing Deliberate Learning Opportunities

11.1.1 Generalizability - Active learnersourcing

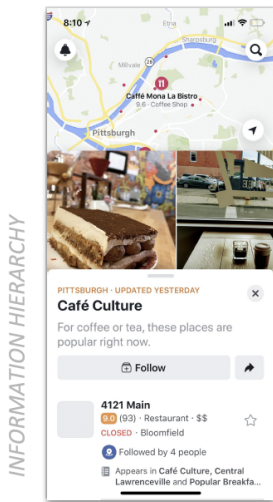
We have observed while applying UpGrade to different courses that the quality of the questions produced is highly correlated with how well structured the data input is. As an example, when the open-ended solutions we passively collect provide nuanced structures, it gives the system more input in creating questions. We suggest these structures can be actively collected by helping instructors design more detailed rubrics when giving open-ended assignments to students. In future work, UpGrade can also help instructors refine learning objectives and coming up with better rubrics.

As defined in [60], passive learnersourcing uses data generated by learners' natural interaction with the learning platform. On the other hand, active learnersourcing prompts learners to engage in specific activities, to provide pedagogical benefits (for the current learner) and to collect useful information (for future learners) at the same time. For my work thus far, I have been exploring passive learnersourcing using student written assignments or short answers as inputs. Future work could explore how active learnersourcing activities can be designed and implemented to bring additional benefits for current and future learners. Two explorations we have done shows promising of leveraging active learnersourcing in creating deliberate practice opportunities. The first example shows that student solutions help instructors refine learning objectives, provide more detailed rubrics and thus enables the system to create finer-grained data inputs. The second example shows that actively giving students structured exercises provides us with input to create versatile multiple-choice questions.

In the course User-centered Research and Evaluation, one learning objective on the method of

#8 Lack of information on location on event page

Individual Event Page



During 0:57 - 1:02, the user was confused on what to do next after selecting an event that interested her

Findings

- The user was confused and unsure about the event because she didn't know where the event was at and for what purpose
- From the recommended list, she selected "Café Culture" as it sounded interesting, but immediately left after 5 seconds as she was unsure of where the event was
 - **Frequency:** Users are likely to experience this issue as a lot of the event pages do not have a clear information hierarchy for each of the hosted events such as addresses, contact, etc.
 - **Impact:** It may be difficult for users to overcome this issue because the only actions that a user can take from here is either to follow the event or share the event to other friends, which isn't solving their needs on figuring out what the address of the event is
 - **Persistence:** Users may be repeatedly bothered by it as it doesn't display the address of the event, instead it uses Facebook's integrated map which doesn't clearly show the users on what next steps to take to get to that event. Another user in [Slide 7] also struggled with this issue.

Recommendations

- **Solution:** Include the address of the publicly hosted event and the point of contact for that event in case of help
- **Trade-offs:** Linking another Facebook user's profile as a point of contact may cause privacy concerns

Based on UAR by Brad A. Myers & Bonnie John

Figure 11.1: Student solution example on the topic of usability findings. The severity of the problem is further broken down into three aspects.

think-aloud protocols is to assess the severity of a usability problem. Through multiple offerings of the course, the learning objective was refined as assessing the severity of a usability problem from three aspects. As shown in 11.1, when evaluating the severity of a usability problem, students were asked to evaluate from three aspects, namely frequency, impact and persistence. With this data input, UpGrade creates multiple-choice questions that exercise this knowledge component of differentiating aspects in assessing a usability issue.

Another important research method students need to learn is to conduct interviews. In the open-ended assignment, students were asked to list the interview questions they will use. Through an active learnersourcing approach, we asked students to select 5 interview questions from the list, revise them and offer an explanation. The activity was done through Google form, as shown in Figure 11.2. The data resulted from this activity was used as sources to create new multiple-choice questions, Figure 11.3 shows an example question created. The pilot study shows promise for using more active learnersourcing practices. On the one hand, it starts by logging student assignment data hierarchically, which saves future efforts to segment the solutions. On the other hand, it helps current students reflect on their solutions and creates additional sources to create questions for future students.

11.1.2 Scalability - Data sharing and reuse

The courses I have worked with have multiple offerings in the past and have existing student solution data to bootstrap the system. For courses without existing resources to start with,

Section 2 of 6

Revision 1

Description (optional)

Original interview question *

Long answer text

Please revise the above interview question *

Long answer text

Explanation of revision *

Long answer text

Figure 11.2: The active learnersourcing practice we piloted in UCRE.

Your classmates are asked to conduct semi-structured interviews with CMU students to understand the transportation around CMU campus. Before they conducted the interview, they drafted an initial list of interview questions and revised these interview questions based on expert feedback or self-reflection. Please find below two versions of an interview question, and select which one is better.

Version 1

Yesterday the temperature was low, did you take CMU shuttle? Why or why not?


Version 2

In what conditions would you prefer CMU shuttle over other transportation methods?

Which one is a better interview question?

Version 1

Version 2



Upgrade Feedback:

Great! You and a previous student both thought the **Version 1** is better. Indeed, **Version 1** is a revision of **Version 2**. **Here's what the student said, do you agree?**

I want to ask the participant about very specific experiences and events instead of letting him/her summarize the experience.

Figure 11.3: An example UpGrade-created question with active learnersourcing.

instructors can either take the active learnersourcing approach to collect student solutions or reuse the materials created from other courses. As examples, two of our modules on research methods were reused across classes at CMU.

To facilitate this data sharing and reuse, privacy issues related to student data may arise. Future work could also explore issues around the ownership of student-generated data, ways to credit students when using their answers as course materials, and ways to share and reuse data in a privacy-preserving way, following school regulations.

11.1.3 Varying inputs, outputs and domains

Getting deliberate practice is critical in developing skill mastery. In our work, from interviews and surveys with instructors, we see that creating enough deliberate practice opportunities is time-consuming and challenging. UpGrade is one attempt at semi-automatically creating deliberate learning opportunities with immediate feedback. The input is past students' open-ended solutions and the output is multiple-choice questions that target higher-order cognitive skills, such as evaluation skills. Future work could explore other sorts of inputs and outputs to support deliberate practice in different domains. For example, forum Q&A, lecture notes, online tutorials (both texts and videos), programs, etc., can all be potential inputs, and besides multiple-choice questions, such systems may create various sorts of practice opportunities with feedback.

My past work has focused on providing deliberate practice opportunities as part of students' take-home assignments. Future work could explore leveraging such techniques to support active in-class instruction. A large body of learning science literature suggests that active learning that encourages students to participate and engage with the content is better than passive lecturing [22, 29]. However, a recent study found that most college STEM instructors still choose traditional passive teaching methods, such as lecturing with little student interaction[125]. An experiment in a college physics course found that active instruction that gives students problems to solve and discuss during class results in more learning than passive instruction, which is having students passively listen to lectures [29]. Future work could employ such content authoring techniques to create problem-solving activities to be embedded during in-class instruction.

11.1.4 Social Learning

In the deliberate practice literature [34], it is often described as an individual mastery learning approach, focusing on repeated practice at the edge of the learner's competency with immediate feedback. The social contexts of where such learning happens are rarely discussed. In the future, building upon my prior work related to collaborative learning and team-based learning [114, 115, 116, 134, 135], I'm also interested in connecting deliberate practice efforts with social learning support, enabling students to engage in deliberate practice together and exchange ideas where appropriate.

11.1.5 Adaptivity

Besides creating deliberate learning opportunities, I'm also excited about adaptively providing deliberate learning opportunities when learners are doing open-ended tasks. For example, when

a student is writing an interview protocol, the system provides deliberate practice opportunities when it identifies the student makes mistakes on certain knowledge components.

11.1.6 Machine Learning Enhanced Quality Control

We see opportunities for leveraging instructor revision history to further speed the revision history. A possible direction would be building predictive models on the quality of questions based on instructor revision history, thus reducing human review time.

We also see opportunities for using more sophisticated natural language processing techniques to filter and compress texts and remove irrelevant information from existing student solutions. Reinforcement learning approaches could also be explored.

11.1.7 Potential Risks and Ways to Mitigate Them

Potential risks exist using systems similar to UpGrade. Crowdsourcing existing solutions may risk broadcasting incorrect information written by one student to more students. Some student solutions may contain irrelevant information to the targeted learning objectives, or that segments of student solutions do not provide enough contexts for future students to fully understand and learn. These may lead to unnecessary extra cognitive load for students trying to learn from these questions and may reduce the expected benefits from such systems.

I summarize three pathways to mitigate the potential risks of deploying such systems. On the one hand, we have seen the importance of expert-involvement in this process. Giving instructors an opportunity to finally revise questions and check information accuracy and comprehensiveness makes sure that the content delivered to students is of high quality. However, the efforts required from experts would also undermine the scalability of such approaches. Techniques to enhance effective human-machine collaboration could reduce instructor efforts. For example, incorporating instructor editing history into the system to prioritize question items likely to be good items, which could reduce the reviewing effort needed. Finally, designing systems that would allow instructors to collect student responses better aligned with the targeted learning objectives would help collect better data input up front, saving subsequent efforts in processing messy data. Detailed discussions of this direction is presented in Section 11.1.1

11.2 Artificial Intelligence in Education

Traditionally, teachers take the most responsibility in providing learning opportunities to students, e.g., giving lectures, offering feedback. However, the efforts required from experts make such learning opportunities less scalable. On the other hand, AI-based technologies have begun to tackle this scaling issue, such as automatic grading (Arrow #1 in Figure 11.4), intelligent tutoring systems that adaptively select questions based on students' knowledge (Arrow #2 in Figure 11.4). But challenges remain. In my work, I show we can take advantage of AI and the open-ended data past students produced to augment teachers' capabilities on content creation (Arrow #3 in Figure 11.4) [139]. I have also explored ways that AI could support team-based learning through signaling opportunities for in-depth discussion (Arrow #4 in Figure 11.4) [131, 136, 145].

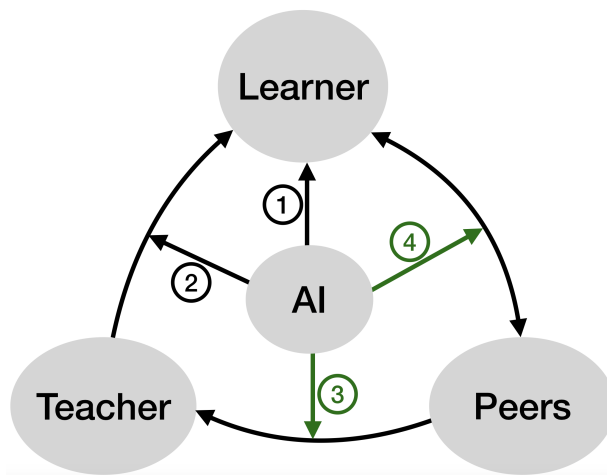


Figure 11.4: Arrows #1 and #2 are existing focuses of AI in education; #3 and #4 indicate my contributions. #1 stands for direct AI support to learners, e.g., automatic grading, chat bots. #2 stands for techniques that intervene during instruction, e.g., intelligent tutoring systems that support adaptive question selection. #3 stands for techniques to support the authoring of content leveraging past student data. #4 stands for techniques to support interaction between peers.

Besides learner-centered systems, future work could explore diverse support links such as Arrow #3 and Arrow #4, as shown in Figure 11.4). And besides learner support systems, I encourage researchers in the field of AIED to explore approaches and find solutions to better support instructors and instructional designers. Notable prior projects in this direction include teacher dashboard that visualizae student performance [149], and real-time teacher augmentation to combine strengths of human and AI instruction in K12 classrooms[54], Overcode[41], a system for visualizing and exploring thousands of programming solutions and enable instructors to offer aggregate feedback to student solutions. In my research, we also see that many of the instructional design and implementation efforts could largely benefit from the advances in AI and technologies, for example, helping instructors define learning objectives, perform cognitive task analysis, design a variety of instructional activities and assessments, etc. My work offers a starting point of helping instructors break things into more manageable steps and provide a bit of scaffolding for them in creating deliberate practice opportunities.

On the one hand, my work demonstrates that advances in computer science have promise in improving education, e.g., natural language understanding techniques can be leveraged for automatic content creation and scalable teaching. On the other hand, I seek to explore how advances in learning technologies could contribute to computer science. For instances, research on human intelligence could inform the design of machine intelligence; Human-AI hybrid methods that can make an impact in practice contribute knowledge around how to better structure and engineer training data as input to algorithms and how to leverage the complementary strengths of human and AI to tackle real-world challenges.

11.3 Human-AI Collaboration

A major takeaway I've gained from my dissertation work is that close collaboration between humans and AI is critical in solving real-world problems as the ones I have shown in generating deliberate practice for higher-order thinking [139], and in repurposing online tutorial videos for software learning [138]. In the example of UpGrade, the complementary strengths of experts, novices and machine intelligence are leveraged in the process of content creation. With instructors alone, it is often hard to write lots of elaborated good examples and wrong answers. With AI alone in prior work, we see that existing question generation techniques for educational purposes only produce fact questions and are not flexible. Here, past peers offer a powerful data source with structured examples that show common errors. With this better data input, machine can auto selectfilter and reorganize examples. Instructors, in the end, will check information accuracy and comprehensiveness. In the example of the workflow categorization and recommendation technique, the partnership between humans the machine is also manifested. Peers who can be users of the software everywhere in the world contribute demonstration videos. We collect these videos (and their command logs) as input and use machine to identify an end user's workflow, expert input is elicited in this process. With this approach, existing online repositories may be repurposed as targeted tutorials for end users.

In this process, I also observed human-AI collaboration issues and opportunities that future work could investigate. As examples, specific to one step in question authoring, selecting plausible distractors, instructors have different preferences in different contexts, such as "I want to use abstract distractors here" or "I want to select more technical answers here". Following human judgements, AI selects answers that match the above criteria, in this case using natural language understanding techniques. In this example, human experts make judgements about which constructs to capture, and subsequently, AI is used to automatically capture these constructs, demonstrating the flexibility of humans and the scalability of AI.

There is an increasing amount of work in HCI to investigate ways that humans and AI could collaborate. As examples, Holstein developed teacher augmentation tools by combining the strengths of human and AI instruction [54], Guo developed systems [47, 48] that combine crowdsourcing systems and computer vision techniques to support visual information access. Building upon these prior efforts in developing human-AI systems, I'd like to explore and define the engaged efforts from different human and machine stakeholders to make such systems work. For example, there could be different types of human stakeholders, including crowd and experts. Some open questions I'd like to answer in my future work include, how to better leverage the crowd power? What types of data can we collect and how can we collect better examples? How to best user experts' time? When and how to elicit expert input? How to foster human-machine communication, e.g., having machines execute human preferences? I'd also like to continue developing and applying these techniques in different domains to solve real-world problems.

11.4 Instruction and Assessment in Higher Education

11.4.1 Scaffold before Open-ended Work

Open-ended assignments are frequently used in colleges and are treated as a major source of practice and learning in many courses. However, often, the quality of students' open-ended work is low because they start working on large, complex, open-ended projects before they are ready. Providing instruction and scaffold beforehand could be a win-win solution, where it saves instructors' subsequent efforts in offering feedback to repeated mistakes and help students produce higher quality content that can go into their portfolios.

11.4.2 Discrepancy between Practice and Theories

Expert blindspot is one reason for the lack of movement from current practice to what learning sciences theories would recommend. We provide evidence that instructors have so-called "expert blind spots", revealed through cases where their beliefs and student performance do not match [98, 99]. Instructor beliefs are important because they will influence the design of curriculum and learning experience of students. In both this and a past case [69], we see experts have good reasons for their beliefs, yet data suggests otherwise and a deeper analysis explains why. 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.

Another reason is perhaps that student perceptions, which could be opposite to the actual learning outcomes, pose further challenges in promoting new pedagogies that lead to cognitive dissonance in classrooms. An experiment in a college physics course found that active instruction that gives students problems to solve and discuss during class results in more learning than passive instruction, which is having students passively listen to lectures [29]. However, surprisingly, students perceived that they learnt more from listening to lectures than doing the active problem-solving. Future research and practice could explore new teaching evaluation models that do not evaluate teaching solely based on student ratings.

11.4.3 Service Design in Higher Education

My work suggests that we need to establish the profession of Learning Experience (LX) designers to develop a curriculum in higher education. College instructors are experts in their domains, but they are not necessarily experts on pedagogy. In many other domains, the design of products to support the workflow of professionals requires expertise from both domain experts and interaction designers, e.g., interaction designers design products to support doctors' decision making [151]. 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 the quality of learning design in higher-education, establishing roles such as learning designers or learning engineers is desirable. Given the increasing number of stakeholders involved in delivering educational service in a higher education context, future work could further investigate

the information flow and find design solutions to help different stakeholders collaborate and deliver the best educational service to students.

11.5 Professional Training and Informal Learning

My dissertation work has focused on formal higher education contexts. However, many concepts presented here would apply to professional training and informal learning contexts outside of schools [115, 141]. As examples, many companies have grown programs to support internal employee training and professional development. Learning is becoming increasingly ubiquitous and easy with the advances in technologies such as web search, online tutorials, conversational agents, mobile apps, etc. In my future work, I'd also like to explore techniques to support professional training and informal learning at scale.

11.6 Generation vs. Evaluation, Creativity

As defined in the KLI framework [70], students gain knowledge components (KCs) through learning events, which can be inferred from performance on assessment events. The KLI framework suggests that kinds of KCs drive instructional event choices. For example, instructional approaches that emphasize recall and spacing of practice may benefit learning of historical facts, vocabulary; whereas instructional approaches that prompt self-explanation in students would be more valuable for learning complex principles, such as Newton's laws. My work attempts to add a new layer to the characteristics of knowledge components (KCs) by comparing the relative "Generation" and "Evaluation" efforts involved in problem-solving. For skills (KCs) that require a non-trivial amount of "Evaluation" efforts, whereas the amount of "Generation" efforts required is minimal, they benefit from "evaluation-type" learning events, such as multiple-choice practice questions that target evaluation goals.

With this prior work as the foundation, there are many open questions remaining. Our work demonstrates a low-effort, experimental comparison technique to help instructional designers find out about the relative "Generation" and "Evaluation" efforts involved in problem-solving. It'll be great if future work could offer theoretical guidance on making the judgement. Other related questions I'd like to explore include the role "Generation" and "Evaluation" efforts play in learning and performance, respectively, and the relationship between "Generation" and "Evaluation" with creativity.

Bibliography

- [1] Canvas. <https://www.canvaslms.com/research-education>, 2019. 7.1
- [2] Eytan Adar, Jaime Teevan, and Susan T Dumais. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 1197–1206, 2008. 9.2.2
- [3] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering*, pages 3–14. IEEE, 1995. 9.4.1
- [4] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993. 9.4.1
- [5] Vincent Aleven, Bruce M McLaren, Jonathan Sewall, and Kenneth R Koedinger. The cognitive tutor authoring tools (ctat): preliminary evaluation of efficiency gains. In *International Conference on Intelligent Tutoring Systems*, pages 61–70. Springer, 2006. 1
- [6] Vincent Aleven, Bruce M McLaren, Jonathan Sewall, and Kenneth R Koedinger. A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, 19(2):105–154, 2009. 3.3
- [7] Vincent Aleven, Jonathan Sewall, Octav Popescu, Martin van Velsen, Sandra Demi, and Brett Leber. Reflecting on twelve years of its authoring tools research with ctat. *Design recommendations for adaptive intelligent tutoring systems*, 3:263–283, 2015. 1
- [8] Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. Ontology-based multiple choice question generation. *KI-Künstliche Intelligenz*, 30(2):183–188, 2016. 1, 2.1
- [9] Susan A Ambrose, Michael W Bridges, Michele DiPietro, Marsha C Lovett, and Marie K Norman. *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons, 2010. 1, 2, 2.5, 3.2, 3.2.2, 3.2.3, 8.1, 8.2.4
- [10] John R Anderson, C Franklin Boyle, Albert T Corbett, and Matthew W Lewis. Cognitive modeling and intelligent tutoring. *Artificial intelligence*, 42(1):7–49, 1990. 2.3.2
- [11] Ewa Pihammar Andersson. The perspective of student nurses and their perceptions of professional nursing during the nurse training program. *Journal of Advanced Nursing*, 18:808–815, 1993. 8.2.2
- [12] Autodesk. About the screencast api and sdk. <https://goo.gl/eNPbw4>, 2017. Accessed: 2017-12-21. 9.9.3
- [13] Nikola Banovic, Tovi Grossman, Justin Matejka, and George Fitzmaurice. Waken: reverse

- engineering usage information and interface structure from software videos. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 83–92, 2012. 9.9.3
- [14] Scott Bateman, Jaime Teevan, and Ryen W White. The search dashboard: how reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1785–1794, 2012. 9.2.1, 9.2.1
- [15] Robert A Bjork. Memory and metamemory considerations in the. *Metacognition: Knowing about knowing*, 185, 1994. 7.2
- [16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 9.4.2
- [17] TAXONOMY MADE EASY BLOOM’S. *Bloom’s taxonomy of educational objectives*. Longman, 1965. 7.3.1
- [18] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006. 5.1.1, 8.4.5
- [19] Sara E Brownell, Jordan V Price, and Lawrence Steinman. Science communication to the general public: why we need to teach undergraduate and graduate students this skill as part of their formal scientific training. *Journal of undergraduate neuroscience education*, 12(1): E6, 2013. 8.1
- [20] Narges Toghian Chaharsoughi, Shahnaz Ahrari, and Shahnaz Alikhah. Comparison the effect of teaching of sbar technique with role play and lecturing on communication skill of nurses. *Journal of caring sciences*, 3(2):141, 2014. 8.1
- [21] Seth Chaiklin. The zone of proximal development in vygotsky’s analysis of learning and instruction. *Vygotsky’s educational theory in cultural context*, 1:39–64, 2003. 2, 2.5, 3.2.2
- [22] Michelene TH Chi and Ruth Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243, 2014. 1, 11.1.3
- [23] Nancy Van Note Chism, Matthew Holley, and Cameron J Harris. 9: Researching the impact of educational development: Basis for informed practice. *To improve the academy*, 31(1): 129–145, 2012. 8.1
- [24] Amy Shannon Cook, Steven P Dow, and Jessica Hammer. Towards designing technology for classroom role-play. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 241–251, 2017. 8.1
- [25] National Research Council et al. *How students learn: History, mathematics, and science in the classroom*. National Academies Press, 2004. 8.1, 8.2.3
- [26] Sandra Crespo. Seeing more than right and wrong answers: Prospective teachers’ interpretations of students’ mathematical work. *Journal of Mathematics Teacher Education*, 3(2): 155–181, 2000. 8.1
- [27] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3): 297–334, 1951. 1, 2, 2.6, 3.2.4, 6.1
- [28] Boris Cuke, Bart Goethals, and Céline Robardet. A new constraint for mining sets in

- sequences. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 317–328. SIAM, 2009. 9.4.1
- [29] Louis Deslauriers, Logan S McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39):19251–19257, 2019. 7.1, 11.1.3, 11.4.2
- [30] Himel Dev and Zhicheng Liu. Identifying frequent user tasks from application logs. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 263–273, 2017. 9.1, 9.2.2, 9.4, 9.4.1, 9.6
- [31] Brent Duckor and Carrie Holmberg. *Mastering formative assessment moves: 7 high-leverage practices to advance student learning*. ASCD, 2017. 8.2.3
- [32] Barbara J Ericson, James D Foley, and Jochen Rick. Evaluating the efficiency and effectiveness of adaptive parsons problems. In *Proceedings of the 2018 ACM Conference on International Computing Education Research*, pages 60–68, 2018. 8.1
- [33] K Anders Ericsson. Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine*, 15(11):988–994, 2008. 2, 2.3.1
- [34] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363, 1993. 1, 2, 2.3.1, 3.2, 3.2.3, 8.2.4, 11.1.4
- [35] K Anders Ericsson et al. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance*, 38:685–705, 2006. 2, 2.3.1
- [36] Education Testing Services (ETS). Reliability and comparability of toefl ibt scores. Technical report. 2.6, 3.2.4
- [37] Steven C Funk and K Laurie Dickson. Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, 38(4):273–277, 2011. 7.3.1
- [38] ERIN MARIE FURTAK and HOWARD M GLASSER. Using formative assessment data for science teaching and learning. 2016. 8.2.3
- [39] Gensim. Gensim library. <https://radimrehurek.com/gensim/>, 2017. Accessed: 2017-07-10. 9.4.2
- [40] David Gerritsen. *A socio-technical approach to feedback and instructional development for teaching assistants*. PhD thesis, Carnegie Mellon University, 2018. 8.1
- [41] Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. Overcode: Visualizing variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(2):1–35, 2015. 11.2
- [42] Steve Graham, Michael Hebert, and Karen R Harris. Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4):523–547, 2015. 1
- [43] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007. 9.2.2

- [44] Pamela Grossman, Christa Compton, Danielle Igra, Matthew Ronfeldt, Emily Shahan, and Peter Williamson. Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9):2055–2100, 2009. 8.1, 8.2.2
- [45] Tovi Grossman and George Fitzmaurice. Toolclips: an investigation of contextual video assistance for functionality understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1515–1524, 2010. 9.2.1
- [46] Tovi Grossman, George Fitzmaurice, and Ramtin Attar. A survey of software learnability: metrics, methodologies and guidelines. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 649–658, 2009. 9.1
- [47] Anhong Guo, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, and Jeffrey P Bigham. Crowd-ai camera sensing in the real world. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–20, 2018. 11.3
- [48] Anhong Guo, Junhan Kong, Michael Rivera, Frank F Xu, and Jeffrey P Bigham. Statelens: A reverse engineering solution for making existing dynamic touchscreens accessible. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 371–385, 2019. 11.3
- [49] Maria Riaz Hamdani. Learning how to be a transformational leader through a skill-building, role-play exercise. *The International Journal of Management Education*, 16(1):26–36, 2018. 8.1
- [50] Deborah Harris. Comparison of 1-, 2-, and 3-parameter irt models. *Educational Measurement: Issues and Practice*, 8(1):35–41, 1989. 2, 2.6, 3.2.4
- [51] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007. 2, 2.5, 3.2.3
- [52] Neil T Heffernan and Cristina Lindquist Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014. 3.3
- [53] Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 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*, pages 458–469. ACM, 2016. 7.1
- [54] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*, pages 154–168. Springer, 2018. 11.2, 11.3
- [55] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 10–19, 2008. 9.2.2, 9.4.2
- [56] Petri Ihantola, Tuukka Ahoniemi, Ville Karavirta, and Otto Seppälä. Review of recent systems for automatic assessment of programming assignments. In *Proceedings of the 10th*

Koli calling international conference on computing education research, pages 86–93, 2010. 1

- [57] Andy Jacob and Kate McGovern. The mirage: Confronting the hard truth about our quest for teacher development. *TNTP*, 2015. 8.1
- [58] Beres Joyner and Louise Young. Teaching medical students using role play: twelve tips for successful role plays. *Medical teacher*, 28(3):225–229, 2006. 8.1
- [59] David A Joyner, Wade Ashby, Liam Irish, Yeeling Lam, Jacob Langston, Isabel Lupiani, Mike Lustig, Paige Pettoruto, Dana Sheahen, Angela Smiley, et al. 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*, pages 399–408. ACM, 2016. 7.1
- [60] Juho Kim. *Learnersourcing : improving learning with collective learner activity*. PhD thesis, Cambridge, MA, USA, 2015. 2, 2.2, 3.2, 3.2.1, 11.1.1
- [61] Juho Kim, Robert C Miller, and Krzysztof Z Gajos. Learnersourcing subgoal labeling to support learning from how-to videos. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 685–690. 2013. 9.2.1
- [62] Juho Kim, Philip J Guo, Carrie J Cai, Shang-Wen Li, Krzysztof Z Gajos, and Robert C Miller. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 563–572, 2014. 9.9.2
- [63] Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 563–572, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3069-5. doi: 10.1145/2642918.2647389. URL <http://doi.acm.org/10.1145/2642918.2647389>. 2, 2.2, 3.1, 3.2, 3.2.1, 10.1
- [64] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 4017–4026, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2556986. URL <http://doi.acm.org/10.1145/2556288.2556986>. 2, 2.2, 3.1, 3.2, 3.2.1, 9.2.1, 9.9.3, 10.1
- [65] Juho Kim, Elena L. Glassman, Andrés Monroy-Hernández, and Meredith Ringel Morris. Rimes: Embedding interactive multimedia exercises in lecture videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1535–1544, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702186. URL <http://doi.acm.org/10.1145/2702123.2702186>. 2, 3.7.3
- [66] Paul A Kirschner and Jeroen Van Merriënboer. Ten steps to complex learning a new approach to instruction and instructional design. 2008. 8.2.2

- [67] Avraham N Kluger and Angelo DeNisi. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254, 1996. 2, 2.5, 3.2.3
- [68] Günther Knoblich, Stellan Ohlsson, Hilde Haider, and Detlef Rhenius. Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, memory, and cognition*, 25(6):1534, 1999. 2, 2.3.2
- [69] Kenneth R Koedinger and Mitchell J Nathan. The real story behind story problems: Effects of representations on quantitative reasoning. *The journal of the learning sciences*, 13(2): 129–164, 2004. 7.1, 7.3.3, 10.2, 11.4.2
- [70] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012. 1, 2, 2.4, 10.2, 11.6
- [71] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 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*, pages 111–120. ACM, 2015. 7.1
- [72] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 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*, pages 111–120. ACM, 2015. 1
- [73] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33:1–33:31, December 2013. ISSN 1073-0516. doi: 10.1145/2505057. URL <http://doi.acm.org/10.1145/2505057>. 2, 2.5, 3.2, 3.2.3, 10.1
- [74] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. Peerstudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 75–84. ACM, 2015. 1, 7.1
- [75] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204, 2020. 1, 2.1, 10.1
- [76] Ben Lafreniere, Tovi Grossman, Justin Matejka, and George Fitzmaurice. Investigating the feasibility of extracting tool demonstrations from in-situ video content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4007–4016, 2014. 9.2.1, 9.9.2
- [77] Benjamin Lafreniere, Tovi Grossman, and George Fitzmaurice. Community enhanced tutorials: improving tutorials with multiple demonstrations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1779–1788, 2013. 9.2.1, 9.2.1
- [78] Magdalene Lampert, Megan Loef Franke, Elham Kazemi, Hala Ghouseini, Angela Chan Turrou, Heather Beasley, Adrian Cunard, and Kathleen Crowe. Keeping it complex: Using

- rehearsals to support novice teacher learning of ambitious teaching. *Journal of teacher education*, 64(3):226–243, 2013. 8.2.2
- [79] Douglas Larkin. Misconceptions about "misconceptions": Preservice secondary science teachers' views on the value and role of student ideas. *Science Education*, 96(5):927–959, 2012. 8.1
- [80] Gilly Leshed, Eben M Haber, Tara Matthews, and Tessa Lau. Coscripiter: automating & sharing how-to knowledge in the enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1719–1728, 2008. 9.2.1
- [81] Arthur Levine. Educating school teachers. *Education Schools Project*, 2006. 8.1, 8.2.2
- [82] Wei Li, Justin Matejka, Tovi Grossman, Joseph A Konstan, and George Fitzmaurice. Design and evaluation of a command recommendation system for software applications. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(2):1–35, 2011. 9.7.1
- [83] Wei Li, Tovi Grossman, and George Fitzmaurice. Cadament: a gamified multiplayer software tutorial system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3369–3378, 2014. 9.2.1, 9.2.1
- [84] SPMF Library. Spmf library. <http://www.philippe-fournier-viger.com/spmf/>, 2017. Accessed: 2017-07-10. 9.6
- [85] Frank Linton and Hans-Peter Schaefer. Recommender systems for learning: Building user and expert models through long-term observation of application use. *User Modeling and User-Adapted Interaction*, 10(2-3):181–208, 2000. 9.2.1, 9.2.1
- [86] Frank Linton, Deborah Joy, Hans-Peter Schaefer, and Andrew Charron. Owl: A recommender system for organization-wide learning. *Educational Technology & Society*, 3(1): 62–76, 2000. 9.1, 9.2.1
- [87] Jeri L Little and Elizabeth Ligon Bjork. Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, 43(1):14–26, 2015. 7.3.1
- [88] Jeri L Little, Elizabeth Ligon Bjork, Robert A Bjork, and Genna Angello. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological science*, 23(11):1337–1344, 2012. 7.3.1
- [89] Ching Liu, Juho Kim, and Hao-Chuan Wang. Conceptscape: Collaborative concept mapping for video learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 387:1–387:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173961. URL <http://doi.acm.org/10.1145/3173574.3173961>. 2, 2.2, 3.1, 3.2, 3.2.1, 10.1
- [90] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 9.9.4
- [91] Sylvain Malacria, Joey Scarr, Andy Cockburn, Carl Gutwin, and Tovi Grossman. Skillometers: reflective widgets that motivate and help users to improve performance. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 321–330, 2013. 9.2.1, 9.2.1
- [92] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes

- in event sequences. *Data mining and knowledge discovery*, 1(3):259–289, 1997. 9.4.1
- [93] Justin Matejka, Wei Li, Tovi Grossman, and George Fitzmaurice. Communitycommands: command recommendations for software applications. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 193–202, 2009. 9.1, 9.2.1, 9.2.1, 9.5.1, 9.7.2, 9.8, 9.9.1
- [94] Justin Matejka, Tovi Grossman, and George Fitzmaurice. Ambient help. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2760, 2011. 9.1, 9.2.1, 9.2.1
- [95] Sarah Michaels and Catherine O’Connor. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue*, pages 347–362, 2015. 8.2.3
- [96] Sonia Moore. *The Stanislavski system: Professional Training of an Actor*. Viking Penguin Inc., 1984. 8.2.2
- [97] Mitchell J Nathan and Kenneth R Koedinger. An investigation of teachers’ beliefs of students’ algebra development. *Cognition and Instruction*, 18(2):209–237, 2000. 7.3.2
- [98] Mitchell J Nathan and Kenneth R Koedinger. Teachers’ and researchers’ beliefs about the development of algebraic reasoning. *Journal for Research in Mathematics Education*, pages 168–190, 2000. 7.1, 7.3.2, 10.2, 11.4.2
- [99] Mitchell J Nathan and Anthony Petrosino. Expert blind spot among preservice teachers. *American educational research journal*, 40(4):905–928, 2003. 7.1, 10.2, 11.4.2
- [100] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 104. Prentice-Hall Englewood Cliffs, NJ, 1972. 2, 2.5
- [101] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 597–606, 2007. 9.2.2
- [102] Stellan Ohlsson. Some principles of intelligent tutoring. *Instructional science*, 14(3-4): 293–326, 1986. 2, 2.3.2
- [103] Stellan Ohlsson. Constraint-based student modeling. In *Student modelling: the key to individualized knowledge-based instruction*, pages 167–189. Springer, 1994. 2, 2.3.2
- [104] Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1):1–4, 2003. 2, 2.5, 3.2.2
- [105] Jaimie Y Park, Neil O’Hare, Rossano Schifanella, Alejandro Jaimes, and Chin-Wan Chung. A large-scale study of user image search behavior on the web. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 985–994, 2015. 9.2.2
- [106] Marco Pennacchiotti and Siva Gurumurthy. Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*, pages 101–102, 2011. 9.4.2

- [107] Adam Perer and Fei Wang. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 153–162, 2014. 9.2.2
- [108] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256, 2009. 9.4.2
- [109] Justin Reich, Yoon Jeon Kim, Kevin Robinson, Dan Roy, and Meredith Thompson. Exploring authenticity and playfulness in teacher practice spaces. 2018. 8.2.2
- [110] Henry L Roediger III and Jeffrey D Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255, 2006. 7.1
- [111] James M Royer, Cheryl A Cisero, and Maria S Carlo. Techniques and procedures for assessing cognitive skills. *Review of Educational Research*, 63(2):201–243, 1993. 7.3.1
- [112] Philip M Sadler, Gerhard Sonnert, Harold P Coyle, Nancy Cook-Smith, and Jaimie L Miller. The influence of teachers’ knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, 50(5):1020–1049, 2013. 8.1
- [113] Mehdi SM Sajjadi, Morteza Alamgir, and Ulrike von Luxburg. 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*, pages 369–378. ACM, 2016. 7.1
- [114] Sreecharan Sankaranarayanan, Cameron Dashti, Chris Bogart, Xu Wang, Majd Sakr, and Carolyn Penstein Rose. When optimal team formation is a choice-self-selection versus intelligent team formation strategies in a large online project-based course. In *International Conference on Artificial Intelligence in Education*, pages 518–531. Springer, 2018. 11.1.4
- [115] Sreecharan Sankaranarayanan, Xu Wang, Cameron Dashti, Haokang An, Clarence Ngoh, Michael Hilton, Majd Sakr, and Carolyn Rose. Online mob programming: Bridging the 21st century workplace and the classroom. 2019. 11.1.4, 11.5
- [116] Sreecharan Sankaranarayanan, Xu Wang, Cameron Dashti, Marshall An, Clarence Ngoh, Michael Hilton, Majd Sakr, and Carolyn Rosé. An intelligent-agent facilitated scaffold for fostering reflection in a team-based project course. In *International Conference on Artificial Intelligence in Education*, pages 252–256. Springer, 2019. 11.1.4
- [117] Richard A Schmidt and Robert A Bjork. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, 3(4):207–218, 1992. 7.3.1
- [118] Daniel L. Schwartz, Jessica M. Tsang, and Kristen P. Blair. *The ABCs of How We Learn*. W. W. Norton & Company, New York, NY, USA, 2016. ISBN 978-0-393-70926-1. 3.2.2, 3.2.3
- [119] Meghan Shaughnessy and Timothy A Boerst. Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student’s thinking. *Journal of Teacher Education*, 69(1):40–55, 2018. 8.2.1

- [120] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 81–88. ACM, 2017. 7.1
- [121] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001. 9.4.2
- [122] Laurie Sleep. The work of steering instruction toward the mathematical point: A decomposition of teaching practice. *American Educational Research Journal*, 49(5):935–970, 2012. 8.2.1
- [123] Megan A Smith and Jeffrey D Karpicke. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, 22(7):784–802, 2014. 7.3.1
- [124] Linfeng Song and Lin Zhao. Question generation from a knowledge base with web exploration. *arXiv preprint arXiv:1610.03807*, 2016. 2.1
- [125] Marilyne Stains, Jordan Harshman, Megan K Barker, Stephanie V Chasteen, Renee Cole, Sue Ellen DeChenne-Peters, MK Eagan, Joan M Esson, Jennifer K Knight, Frank A Laski, et al. Anatomy of stem teaching in north american universities. *Science*, 359(6383):1468–1470, 2018. 11.1.3
- [126] Ann Stes, Liesje Coertjens, and Peter Van Petegem. Instructional development for teachers in higher education: Impact on teaching approach. *Higher education*, 60(2):187–204, 2010. 8.1
- [127] Brenda Sugrue. A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement: Issues and Practice*, 14(3):29–35, 1995. 7.7
- [128] Brenda Sugrue, Noreen Webb, and Jonah Schlackman. The interchangeability of assessment methods in science. cse technical report 474. 1998. 7.3.1, 7.3.3
- [129] John Sweller. The worked example effect and human cognition. *Learning and instruction*, 2006. 2, 2.5, 3.2, 3.2.2
- [130] Mohsen Tavakol and Reg Dennick. Making sense of cronbach’s alpha. *International journal of medical education*, 2:53, 2011. 2.6, 3.2.4, 6.3
- [131] Gaurav Singh Tomar, Sreecharan Sankaranarayanan, Xu Wang, and Carolyn P Rose. Coordinating collaborative chat in massive open online courses. Singapore: International Society of the Learning Sciences, 2016. 11.2
- [132] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980. 1
- [133] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 225–236, 2016. 9.2.2
- [134] Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé. Investigating how student’s cognitive behavior in mooc discussion forums affect learning gains. *International Educational Data Mining Society*, 2015. 11.1.4
- [135] Xu Wang, Miaomiao Wen, and Carolyn P Rosé. Towards triggering higher-order thinking

- behaviors in moocs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 398–407. ACM, 2016. 11.1.4
- [136] Xu Wang, Miaomiao Wen, and Carolyn Rosé. Contrasting explicit and implicit support for transactive exchange in team oriented project based learning. Philadelphia, PA: International Society of the Learning Sciences., 2017. 11.2
- [137] Xu Wang, Yali Chen, Amanda Godley, and Carolyn Rosé. Public peer review motivates higher quality feedback. In *International Conference of the Learning Sciences*, 2018. 10.1
- [138] Xu Wang, Benjamin Lafreniere, and Tovi Grossman. Leveraging community-generated videos and command logs to classify and recommend software workflows. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018. 1, 11.3
- [139] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale, L@S '19*, pages 17:1–17:10, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6804-9. doi: 10.1145/3330430.3333614. URL <http://doi.acm.org/10.1145/3330430.3333614>. 1, 7.7, 8.1, 8.2.4, 11.2, 11.3
- [140] Xu Wang, Carolyn P Rosé, and Kenneth R Koedinger. To use multiple-choice questions or not? instructor beliefs and student data give different answers. 2019 (in submission). 1
- [141] Xu Wang, Meredith Thompson, Kexin Yang, Dan Roy, Kenneth Koedinger, Carolyn Rose, and Justin Reich. Practice-based teacher education with elk: A role-playing simulation for eliciting learner knowledge (in submission). 2020. 1, 11.5
- [142] Noreen M Webb, Jonah Schlackman, and Brenda Sugrue. The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education*, 13 (3):277–301, 2000. 7.3.3
- [143] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, pages 405–416, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675219. URL <http://doi.acm.org/10.1145/2675133.2675219>. 2, 2.2, 3.1, 3.2, 3.2.1, 10.1
- [144] Miaomiao Wen and Carolyn Penstein Rosé. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1983–1986, 2014. 9.2.2, 9.4.2
- [145] Miaomiao Wen, Keith Maki, Xu Wang, Steven P Dow, James Herbsleb, and Carolyn Rose. Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses. *International Educational Data Mining Society*, 2016. 11.2
- [146] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. Axis: Generating explanations at scale with

- learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale, L@S '16*, pages 379–388, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3726-7. doi: 10.1145/2876034.2876042. URL <http://doi.acm.org/10.1145/2876034.2876042>. 2, 2.2, 2.6, 3.1, 3.2, 3.2.1, 3.2.4, 10.1
- [147] Angelica Willis, Glenn Davis, Sherry Ruan, Lakshmi Manoharan, James Landay, and Emma Brunskill. Key phrase extraction for generating educational question-answer pairs. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale, L@S '19*, pages 20:1–20:10, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6804-9. doi: 10.1145/3330430.3333636. URL <http://doi.acm.org/10.1145/3330430.3333636>. 7.7
- [148] Mark Wilson and Paul De Boeck. Descriptive and explanatory item response models. In *Explanatory item response models*, pages 43–74. Springer, 2004. 2.6, 3.2.4
- [149] Françeska Xhakaj, Vincent Aleven, and Bruce M McLaren. Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. In *European Conference on Technology Enhanced Learning*, pages 315–329. Springer, 2017. 11.2
- [150] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456, 2013. 9.1, 9.4.2
- [151] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 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*, pages 4477–4488. ACM, 2016. 7.1, 11.4.3
- [152] Nesra Yannier, Kenneth R Koedinger, and Scott E Hudson. 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*, pages 1045–1054. ACM, 2015. 1, 8.2.4, 10.2
- [153] Mi Yu and Kyung ja Kang. Effectiveness of a role-play simulation program involving the sbar technique: A quasi-experimental study. *Nurse Education Today*, 53:41–47, 2017. 8.1
- [154] Ken Zeichner. The turn once again toward practice-based teacher education. *Journal of teacher education*, 63(5):376–382, 2012. 8.1
- [155] Yi Zhang, Jamie Callan, and Thomas Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88, 2002. 9.4.2