

DesignQuizzer: A Community-Powered Conversational Agent for Learning Visual Design

ZHENHUI PENG*, Sun Yat-sen University, China

QIAOYI CHEN, Sun Yat-sen University, China

ZHIYU SHEN, Sun Yat-sen University, China

XIAOJUAN MA, The Hong Kong University of Science and Technology, China

ANTTI OULASVIRTA, Aalto University, Finland

Online design communities, where members exchange free-form views on others' designs, offer a space for beginners to learn visual design. However, the content of these communities is often unorganized for learners, containing many redundancies and irrelevant comments. In this paper, we propose a computational approach for leveraging online design communities to run a conversational agent that assists informal learning of visual elements (e.g., color and space). Our method extracts critiques, suggestions, and rationales on visual elements from comments. We present DesignQuizzer, which asks questions about visual design in UI examples and provides structured comment summaries. Two user studies demonstrate the engagement and usefulness of DesignQuizzer compared with the baseline (reading reddit.com/r/UI_design). We also showcase how effectively novices can apply what they learn with DesignQuizzer in a design critique task and a visual design task. We discuss how to use our approach with other communities and offer design considerations for community-powered learning support tools.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools; Empirical studies in HCI**.

Additional Key Words and Phrases: Online communities, visual design, comment processing, informal learning

ACM Reference Format:

Zhenhui Peng, Qiaoyi Chen, Zhiyu Shen, Xiaojuan Ma, and Antti Oulasvirta. 2024. DesignQuizzer: A Community-Powered Conversational Agent for Learning Visual Design . *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 44 (April 2024), 40 pages. <https://doi.org/10.1145/3637321>

1 INTRODUCTION

Online design communities, e.g., Reddit r/UI_design (Figure 1), offer a public space for individuals to learn about design [5, 7, 16, 32, 56, 63, 107]. For instance, the shared design examples and constructive feedback can help novices learn how a key design principle (e.g., the color theme of a webpage) is enacted or violated in practice [4, 31, 36]. Professional designers also participate in design communities to seek inspiring samples and share opinions [43, 49]. This makes these communities more appealing for novices, because it offers a way to learn through peripheral

*Corresponding author.

Authors' addresses: Zhenhui Peng, pengzh29@mail.sysu.edu.cn, Sun Yat-sen University, Zhuhai, China; Qiaoyi Chen, chengy99@mail2.sysu.edu.cn, Sun Yat-sen University, Zhuhai, China; Zhiyu Shen, shenzhy23@mail2.sysu.edu.cn, Sun Yat-sen University, Zhuhai, China; Xiaojuan Ma, mxj@cse.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China; Antti Oulasvirta, antti.oulasvirta@aalto.fi, Aalto University, Helsinki, Finland.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2024/4-ART44 \$15.00

<https://doi.org/10.1145/3637321>

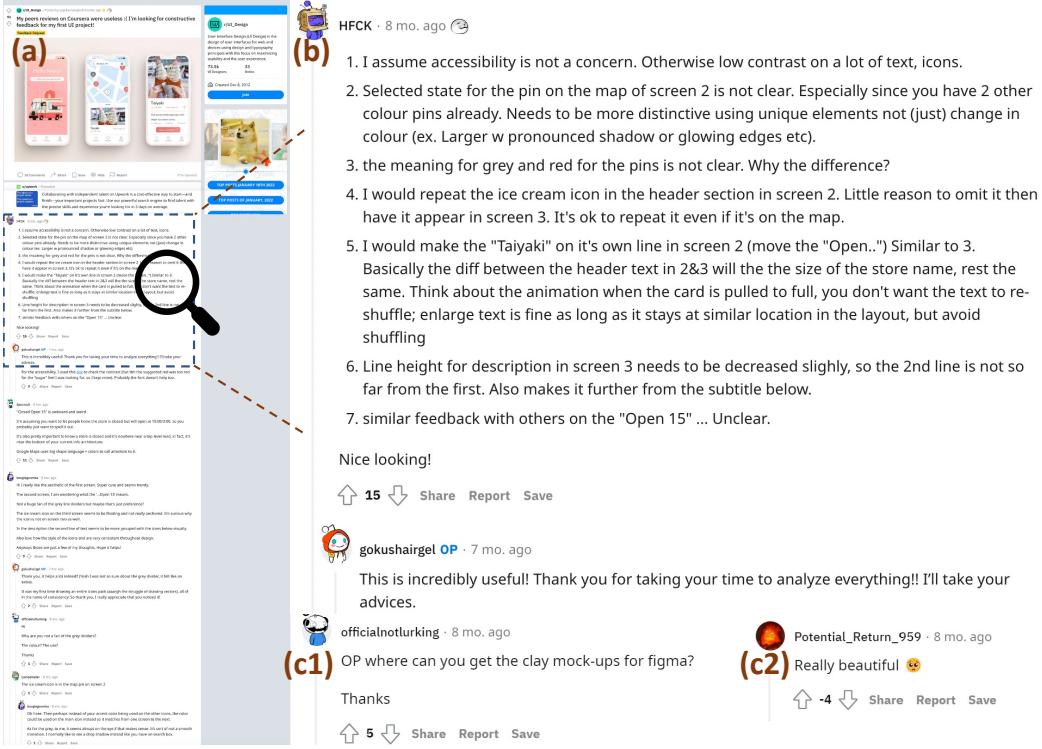


Fig. 1. Online design communities like Reddit r/UI_Design offer design examples and comments as learning resources for novices. For instance, under a UI feedback-request post (part a), there are comments (e.g., b) providing meaningful feedback on the example's visual design. However, many comments are general (e.g., c1, c2) and often unstructured from learners' aspects regarding the types of feedback, UI components, and visual elements. In this paper, we develop computational approaches for extracting and structuring meaningful feedback on UI design examples, based on which we present DesignQuizzler to assist beginners in learning visual design.

participation. This paper focuses on facilitating novices to learn the elementary design principles governing visual elements, such as shape, color, space, form, line, value, and texture [68], from design examples and comments in online communities. Our focus is motivated, on the one hand, by the fact that these elements are a good starting point for novices to learn about visual design, because they are important, visible, and concrete [19]. On the other hand, the commentary in design communities commonly talks about the visual elements of the designs [52], making it possible to build computational models to support learning at scale in informal settings outside classrooms. Our primary targeted user groups are novice designers who only take the user-generated content in the communities as learning materials (merely readers) but do not participate in the content creation. This target is different from previous work on design support tools [42, 46] or on learning in communities by directly interacting with others via posting and commenting [12, 13, 50]. By exploring the materials in design communities, novices are expected to develop their domain skills in visual design, which could be reflected by their critiques of others' designs using learned knowledge and applications of the knowledge to their visual design [14, 22, 23, 36, 84, 106].

However, it is challenging for beginners to effectively and engagingly leverage the examples and comments in online communities to learn visual design. For one thing, peer comments are of varying quality [3, 67, 101, 102] and are typically unstructured for learning purposes (Figure 1), making it time-consuming for users to find helpful content. Existing works have predicted the helpfulness of a comment for the poster of designs based on its language features [51] and qualitatively categorized the content of the feedback [52]. Nevertheless, few of them have computationally structured the comments regarding the types of meaningful feedback (e.g., critique, suggestion [52]) and visual elements that learners are interested in.

For another, people tend to passively read the posts and comments, which could be less engaging and effective than interactively conversing with a conversational agent (CA) for learning [18]. According to the ICAP framework [18] that details the potential benefits of different learning activities (Interactive, Constructive, Active, and Passive), CAs can act as human partners to help users construct their own understandings (constructive engagement) or co-develop explanations with others (interactive engagement). These forms of engagement are more conducive to learning than just reading materials (passive engagement) or manipulating them (e.g., underlining, copying) without new ideas that go beyond the information given (active engagement) [18]. CAs have been demonstrated to be engaging and effective for learning tasks like factual knowledge [82], programming concepts [98], and argumentation skills [93]. Many of them adopt the quiz-based interaction design that prompts users with a question or a task, requires them to answer or finish it, and provides feedback on their performance [77, 82, 93, 98]. Nevertheless, little work has looked into the design, effectiveness, and user experience of a community-powered conversational agent for learning visual design. Unlike previous knowledge acquisition tasks in which there is usually a correct answer for each question, the creative nature of the visual design makes it challenging to prepare quizzes and corrective feedback. Moreover, previous educational CAs usually require experts to curate and label learning materials from official sources, which would be less scalable compared to computationally leveraging rich resources from online communities. Our vision is a computer-supported mode of informal learning where novices can pick up knowledge points by efficiently exploring their interesting content and get engaged in this process by interacting with a conversational agent that generates meaningful learning materials from online communities.

To this end, we develop a computational workflow to structure comments from online design communities into categories of meaningful feedback, based on which we present DesignQuizzer, which facilitates users to learn visual elements by prompting relevant quizzes around given design examples. We present methods that summarize the meaningful feedback from the original comments, classify the summarized feedback sentences into “critique”, “suggestion”, and “rationale”, recognize the keywords about visual elements and UI components (e.g., button, card) in the feedback sentence, and cluster these keywords into more abstract design concepts. We apply this workflow to the comments of UI feedback-request posts in Reddit r/UI_Design, creating a quiz pool that supports DesignQuizzer to ask questions on a given design example’s visual elements, highlight the classified sentences and keywords in the comment summary, and retrieve questions relevant to users’ interesting visual elements or UI components.

We first conduct a within-subjects experiment I with 24 participants to evaluate DesignQuizzer’s user experience and usefulness compared to the r/UI_Design web page baseline. The results suggest that DesignQuizzer improved participants’ efficiency in exploring helpful UI examples and comments and engagement in the visual design learning process. Participants produced more points about the targeted visual elements (e.g., color and typography) with Designquizzer than with the baseline tool when criticizing others’ designs. Nevertheless, due to the potential learning effect on novices’ performance in the design tasks, the experiment I did not evaluate DesignQuizzer’s impact on the application of the learned knowledge to their designs, which is also an important visual

design skill. We then conduct a between-subjects experiment II with another 28 participants. The results suggest that participants in both groups of DesignQuizzer and the r/UI_Design baseline could use the learned knowledge about color and typography to design more consistent, distinct, and intentional UIs in the post-test visual design task compared to the pre-test task. In both experiments, participants felt that DesignQuizzer was significantly more useful and easier to use. They favored its structured comments that ease the reading workload and its quiz-like interaction that encourages active thinking. Yet, they suggested that DesignQuizzer should further incorporate professional knowledge from external resources and enhance users' sense of community. We further discuss ways to apply our approach to other communities for computer-supported learning activities.

This work contributes to CSCW communities from three aspects. First, we develop a computational workflow that makes use of the large-scale comment data in online communities for learning purposes. This workflow can recognize the critique, suggestion, and rationale of the visual elements and UI components in the comments of UI design feedback-request posts. Second, we present an educational application DesignQuizzer for facilitating informal learning activities in online communities. Third, we extend empirical understandings of how people can learn with online user-generated content via two user studies and provide insights into future community-powered learning support tools.

2 RELATED WORK

Our work is built on previous studies on learning in online communities, online design communities, computational methods for modeling comments, and conversational agents in educational domains.

2.1 Learning in Online Communities

In the last two decades, we have witnessed the proliferation of online communities that seek to promote informal learning through unstructured activities and social interactions with others. Cheng et al. described three distinct categories of the learning outcomes common to informal learning communities: development of domain skills, development of community identity, and development of community practices [13]. Developing domain skills refers to the acquisition of knowledge necessary for a person to carry out the core tasks, such as computer programming [13, 81], fan fiction writing [8], and encyclopedia article editing [40]. The second type of outcome involves the development of identity as a member of the community like developing relationships, affinities, and a sense of belonging [13], while the third type means assimilating "cultural artifacts, norms, and values" developed in the community over time [6]. Our work aims to facilitate novices to a discipline in developing their domain skills by exploring the content in online communities. As suggested by previous quantitative studies on learning in communities, the learning of domain knowledge can be captured by the size of the learners' repertoire in terms of the number of types of concepts users can enact after the learning sessions, e.g., computational thinking concepts demonstrated in their projects [14, 22, 23, 84, 106]. Following these works, we measure learners' developed skills by how many enacted or violated principles about the target visual elements they can recognize on an unseen UI example after the learning session. Apart from this measure, we also examine learners' design skills and compare their performance in a visual design task before and after the learning session in experiment II.

Previous work mostly seeks to understand and support community members' learning that happens through sharing their creative artifacts [22, 35, 67, 89] like design mock-ups [15, 16] and/or through social interactions around these artifacts like commenting and critiquing [15, 86, 90]. For example, social computing scholars have documented the way users work together to learn writing and web development skills in fan communities such as FanFiction.net and Archive of Our Own (AO3) [9, 33] as well as programming skills in creative coding communities such as Scratch

[12, 13, 81]. In [13], Cheng et al. further provided a quantitative analysis of legitimate peripheral participation (e.g., engagement with practice proxies and feedback exchange) and learning outcomes in a programming community, suggesting that users' early participation in an online community is associated with long-term learning outcomes [13]. However, little work has explored how to facilitate those who explore others' artifacts and interactions in the communities without direct participation in the posting and commenting activities. Our study fills this gap by supporting these "mere readers" in leveraging community-generated resources for learning purposes.

2.2 Online Design Communities

Online design communities have been found to be beneficial for learners in design. For example, people can post their creative works in these communities to get more timely and "more equal, collaborative, and interactive" [50] critiques for improvement compared to traditional classrooms and workplaces. In fact, many instructors have adopted online peer feedback to support students' learning as their professional and personalized feedback are not always available [53, 74]. Similar to many other communities in domains like creative writing [9] and programming [12], online design communities also support interest-driven learners with shared design examples and associated feedback to gain knowledge [45]. For instance, the members' critiques on the design examples and suggestions can help learners discern what concepts have been executed effectively in the examples [52, 70].

However, it is non-trivial for interest-driven learners to find helpful and needed content in online design communities efficiently and get engaged in this informal learning activity. On the one hand, the peer feedback online is often not as meaningful as that from experts [34, 44, 109]. Although the communities have enabled sorting comments and posts based on keywords and ranking mechanisms (e.g., "Best", "Top", "New", "Controversial", and "Q&A"), novices who lack domain knowledge still find it hard to specify relevant keywords to search [107]. Besides, the comments after sorting may still be unstructured and contain redundant information [64, 102], which hinders learners from locating the content they are interested in. On the other hand, the organization of posts and comments in online design communities tends to offer people a passive-receiving learning experience. According to the ICAP framework [18], the learners' engagement with learning materials can range "from passive to active to constructive to interactive" [18] and would result in an improved learning outcome. Reading entire text passages (e.g., the comments of the design examples) silently without active highlighting and note-taking is a typical passive-receiving activity [18], while asking and answering comprehension questions with a partner is an interactive learning activity [18]. Conversational agents can play the roles of such partners to promote interactive engagement in learning activities and have been demonstrated engaging and effective in learning tasks like argumentative writing [93] and factual knowledge [82].

In all, our research is motivated by the benefit of online design communities for novices' learning tasks, and we explore an effective and engaging way for them to leverage the shared design examples and comments to learn visual design.

2.3 Computational Methods for Modeling Comments

To support efficient comment exploration, existing HCI work has explored a variety of computational techniques to help users filter, structure, and digest the comments in online communities. For example, researchers on community-based question-answering platforms exploit text summarization techniques to extract concise takeaways from threads [47, 87]. Literature about online health communities widely applies document/sentence classification methods, e.g., random forests, linear regressors, and neural networks, to predict the satisfaction level and the amount of sought/received support expressed in members' text [75, 78, 85, 94, 104, 105]. As for the design domain, Krause et

al. rated the helpfulness of the feedback from 176 online providers on students' design solutions and extract a set of natural language features (e.g., specificity, sentiment, etc.) that correlated with the ratings [51]. Yen et al. developed an interactive visualization tool named Decipher that helps designers group the received feedback based on sentiment, keywords (e.g., typography, color) and their interpretation (e.g., fix, keep in mind, need clarification) [108]. However, little work has tried to computationally structure the comments in online design communities from learners' perspectives, e.g., about the mentioned UI components and visual elements of their interests. While there are qualitative findings about what types of design feedback (e.g., the specific, actionable, and justified ones) are perceived as helpful and of a high quality [70, 102], it is under-investigated how to model these types. In this paper, we complement these previous works with a computational approach to model comments regarding the critiques, suggestions, and rationales on visual elements and UI components in design examples shared in online design communities.

2.4 Conversational Agents in Educational Domains

To engage users in their learning tasks, HCI communities have designed and developed conversational agents that converse with users in diverse knowledge domains [25, 71, 96]. For example, AutoTutor has been used to teach college students in computer literacy and critical thinking [71]. AutoTutor provides explanations, feedback, scaffolding, deep reasoning questions, and subject content in online courses, and multiple studies have demonstrated its effectiveness in improving learning gains [71]. Similarly, Wambsganss et al. designed ArgueTutor that judges the argumentative writing performance of users' essays and suggests how to improve [93]. As for the interactive strategies, these agents commonly adopt the quiz-like design by asking questions and giving feedback. For example, Ruan et al. created QuizBot, an interactive agent that asks questions and provides corrective feedback to users' answers in learning factual knowledge about science, safety, and English vocabulary [82]. They showed that QuizBot engaged users better in the learning process than the traditional flashcard tool, and users preferred the bot strongly for casual learning [82]. Winkler et al. developed Sara which acts like a teacher to ask students questions during an online video lecture about programming [98]. They demonstrated in a lab experiment that Sara could significantly improve learning gains compared to the without-Sara condition [98]. Peng et al. proposed a CReBot that prompts critical thinking questions and showed its engagement and usefulness for routine paper readers in their critical paper reading process when compared to a static question list [77]. To power these agents, previous work normally collects quiz questions, prepares and labels answers, and then builds computational models (e.g., classifiers) to provide adaptive feedback to learners [82, 93, 98]. They would require additional human effort to extend the quiz pool when incorporating new learning materials, e.g., factual knowledge about medicine (QuizBot [82]), argumentative writings about scientific papers (Arguetutor [93]), or another online video lecture (Sara [98]).

In this paper, we explore the possibility of a conversational agent helping novices learn visual design. Similar to previous educational agents [77, 82, 98], our proposed DesignQuizzer uses questions to drive the learning process because they are generally effective in encouraging thinking [57, 62, 65, 88]. Unlike previous agents that could be limited to a small number of labeled learning materials, we seek to power DesignQuizzer with computer-generated meaningful materials from online communities. In all, to the best of our knowledge, our work is the first to probe the design, effectiveness, and user experience of a community-powered conversational agent for learning visual design.

3 A COMPUTATIONAL WORKFLOW FOR STRUCTURING COMMENTS IN DESIGN COMMUNITIES

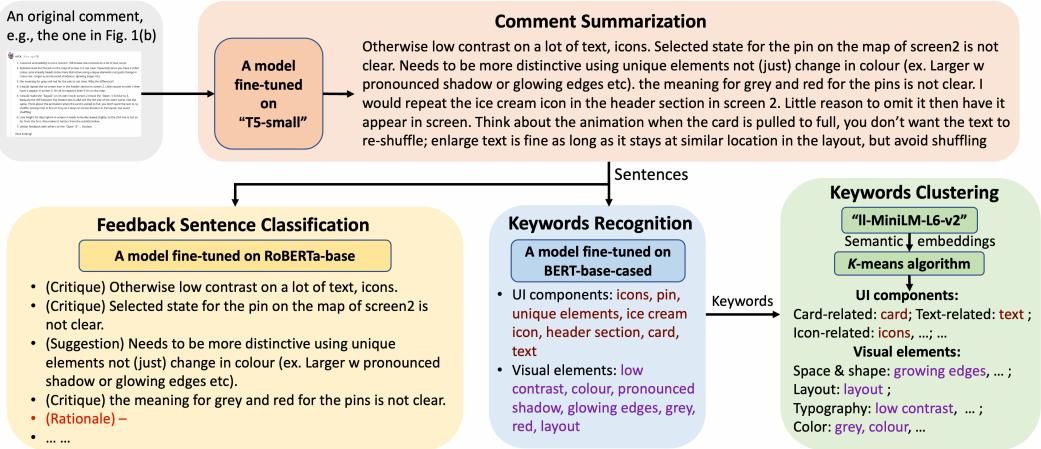


Fig. 2. Computational workflow that structures the information of critique, suggestion, rationale, UI component and visual element keywords, and the clusters of these keywords from the comments under feedback-request posts in online design communities.

In this section, we present our computational workload (Figure 2) to extract useful content from comments as “design quizzes” to facilitate novices in visual design learning. We outline four desirable properties of such quizzers. First, they should include meaningful feedback (specific [51, 109] - critique, actionable [27, 54] - suggestion, and justified [70] - rationale) on the designs. This yields the need for a text summarization model. Second, the quizzes should contain classified feedback sentences that help learners understand how well the design example performs (critique), how to improve it (suggestion), and why giving this critique or suggestion (rationale) [65]. Third, they should be customized based on learners’ interests, e.g., on visual elements or UI components, which requires a token classification method to identify the related keywords. Fourth, they should support the learning session with a learning focus, e.g., on the color-related visual elements. Therefore, we need to cluster the identified keywords about visual elements into higher-level concept groups. In all, our computational approach can extract the following terms from comments of UI feedback-request posts:

- **UI component:** keywords that describe the building blocks for creating the UI examples, e.g., bar, backdrop, button, card, menus, sliders, login page, the title of the post, etc [37].
- **Visual element:** keywords that are the basic units of any visual design which form its structure and convey visual messages, e.g., color, space, shape, font, and size, as well as keywords that describe such units, e.g., aligned, alignment, white, black, hierarchy, readability [68].
- **Critique:** sentences that tell the advantages or drawbacks of specific UI components or visual elements in the design examples [51, 109]. For example, “the locking system seems great and very protective of the space” is a critique of our interests, while “good work” is not.
- **Suggestion:** sentences that provide actionable recommendations about specific UI or visual elements [27, 54], e.g., “You may consider having the shape resemble that of a cup”.
- **Rationale:** sentences that justify the critique or suggestion [70], e.g., “I’d think the title of the post is the most important part of it”.

3.1 Data Collection and labeling

3.1.1 Data collection. To support the supervised learning tasks in the workflow, we used the PushShift api [79] to collect all posts with a “Feedback Request” flair and associated comments created between 2019.5.30 and 2022.5.30 in Reddit r/UI_Design (Figure 1). We chose the r/UI_Design for three reasons. First, it is a typical online design community in Reddit, offering a free place for designers, especially those with little chance to receive feedback from private feedback exchange groups or professional critique services [20], to discuss artworks and exchange critiques [17, 20]. Second, the shared artworks and exchanged critiques in r/UI_Design focus on the design of graphic user interfaces for websites and mobile devices, which matches our focus on facilitating the learning of visual design. Third, founded in 2012, r/UI_Design has over 119K members until May 2023, ranking top 1% by size in Reddit. In other words, r/UI_Design is a representative online design community with valuable resources for learning visual design and developing our computational workflow.

To mitigate the data labeling workload brought by the low-quality comments, we took one more step to increase the likelihood that meaningful feedback exists in our intended labeled comment dataset. Specifically, we used the words “thank”, “great”, “good”, ‘agree”, “advice”, and “suggest” (adapted from [78]) to filter the original posters’ replies to the comments, as previous work suggests that higher-quality comments tend to evoke the original posters to reply and express satisfaction [78]. This step resulted in 2250 comment-reply pairs. To validate if the filtered replies express posters’ satisfaction and if the comments contain meaningful feedback, three authors of this paper independently assigned binary labels to 100 comment-reply pairs and resolved the disagreement via majority voting. The result indicates that 96 out of the 100 replies do express gratefulness (ICC = 0.851) and 72 comments contain meaningful feedback (ICC = 0.751). Thus, we use these 2250 comments under the feedback-seeking posts as the source for the following fine-grained data labeling step and qualitative analyses of the trained computational models.

3.1.2 Data labeling. We randomly sampled 200 of the 2250 comments for the data labeling task. We recruited two annotators (females, ages: 20, 20) from a local university to work on a sequence labeling project in doccano [26]. For each comment, the annotators needed to 1) assign “critique”, “suggestion”, or “rationale” to the appropriate sentence and 2) tag “UI component” or “visual element” to the related keywords¹. We invited a master student (male, age: 24) who majors in industrial design engineering and has completed over ten UI design projects to train our annotators for this labeling task. The training sessions started by asking the annotators to familiarize themselves with the comments in r/UI_Design for one hour and learn visual elements and UI components in Material Design [37] for three days. Next, the master student showcased how he labeled 10 comments, following which our two annotators independently labeled another 20 comments. They then met with the master student to discuss and refine the code book, after which they applied it to another 30 comments. Next, they discussed with the master student again and slightly revised the code book. Finally, they applied the code book for each label (introduced above) to all 200 comments. To evaluate the level of agreement between two annotators, we adopted the ROUGE criteria for the labeled sentences [58]². Specifically, we concatenated the labeled critique sentences separately for each comment to form a paragraph and calculated the ROUGE scores between the paragraphs from two annotators. We did the same for the labeled suggestion and rationale sentences. The results (Table 1) showed that the ROUGE scores reached around 0.8 in all common metrics, indicating a good level of agreement. As for the labeled keywords, there are 80 (8.8%) and 59 (10.6%) times

¹The sentence can be the parts separated by “;”, “and”, “but”, etc. The keywords can consist of one or multiple successive words.

²A common NLP practice to measure the text difference. A higher ROUGE score ranging from 0 to 1 indicates that the texts are more similar.

Table 1. The average ROUGE scores of labeled comment sentences (in percentage).

Sentence Type	ROUGE-1	ROUGE-2	ROUGE-L
Critique	80.42%	79.84%	80.38%
Suggestion	78.22%	77.47%	78.22%
Rationale	90.24%	90.06%	90.24%

of disagreement on the annotated UI components and visual elements. The disagreements were resolved by discussions among the two annotators, the master student, and the first author who coordinated and participated in the training process. In the end, we have 307 critique sentences, 366 suggestions, 155 rationales, 873 keywords about UI components, and 583 visual elements from 173 comments, while the rest 27 do not have any aforementioned labels.

3.2 Comment Summarization

To extract meaningful feedback from the comments, we developed a text summarization model that takes a comment as input and outputs sentences about critiques, suggestions, and rationales. Following [59], we first took an abstractive summarization approach that could better capture the overall meanings of the comment and then fuzzy-matched the output sentences to those in the original comment to help users locate the information they need. For each labeled comment, the sentences with labels “critique”, “suggestion” and “rationale” are concatenated as the ground truth of the output. For the comments ($N = 23$) without meaningful feedback, the ground truth is an empty string “”. We followed the tutorial from HuggingFace on how to fine-tune transformers for downstream summarization task [29], with high regularization and length penalty to ensure that the output sentences used a similar vocabulary to the original text. We experimented with the pre-trained “sshleifer/distilbart-cnn-12-6” model [99] and “T5-small” model [80] using a 7-3 training-testing split method and found that the later one achieves better performance on the test set (e.g., 46.23% vs. 59.90% regarding ROUGE-1 F1 score [58]). Given the small size of our labeled dataset, we used all 200 comments to fine-tune our final “T5-small” summarization model. These summarized texts achieved a high ROUGE-1 / ROUGE-2 / ROUGE-L / ROUGE-L-sum [58] similarity score of 80.98% / 81.53% / 80.98% / 82.29% compared to the ground truth. To provide a concrete context for the summarized texts, we fuzzy-matched each output sentence to the best correspondent sentence and concatenated them as the final output of the comment summarization model.

Table 2 shows the comment summarization results of eight randomly sampled unlabeled comments. The summaries of comments No.1-4 contain meaningful feedback about critique, suggestion, and rationale on specific UI components (e.g., “Apply Filters” in No.2) or visual elements (e.g., line height in No.4). Our model did not capture the general feedback like “That is nice, like it a lot actually”. However, it missed some sentences that could be meaningful (e.g., “At the moment it kinda blends with the background” in No.6). Therefore, we consider that our model’s summary may not cover all the helpful information in the comment, but all the output sentences are meaningful feedback on the UI designs.

3.3 Feedback Sentence Classification

Our feedback sentence classifier inputs a sentence from the comment summary and outputs a label “critique”, “suggestion”, or “rationale”. Our dataset for this supervised multi-class classification task consists of 307 critique, 366 suggestion, and 155 rationale sentences. Using the 8-2 hold-out method, we had 661 sentences for training and the rest 167 for testing. Given the small sample size, we decided to exploit the strength of pre-trained language models to boost the classifier’s performance. We experimented with various BERT-style models including BERT [24], DistilBERT [83], ALBERT

Table 2. Examples of the unlabeled comment under feedback-seeking posts and corresponding summary from our model.

No.	Original comment and its summary (marked in bold) from our model
1	My one suggestion would be to divide the card number, expiry and cvc into visually separate input boxes. Or even just add a dividing line.
2	Not much to complain about here, looking good. I might suggest that the tab bar icons have a lot of varying complexity, maybe they could be more unified or simplified. Also really really nit picky, but you have a pattern of capitalizing the first letter of a statement, except for “Apply Filters”. Maybe just “Apply” would work if “Apply filters” looked weird in a CTA. Thanks for sharing!
3	This is nice, like it a lot actually. Only thing I’d say is maybe the icons could be more unified & there’s possibly not enough difference between the search field and the tiles below. Other than that it’s great. Good job mate!
4	I’d bump the line height of the paragraph a bit.
5	They sorta feel odd.
6	Maybe take the highlight of the background down, make it darker. Thus your product will be the main focus of the page. At the moment it kinda blends with the background.
7	I’m curious as to why the train trip isn’t listed at the top with the hotel and flight?
8	Really nice take on the stepper component.

Table 3. The performance of experimented models for feedback sentence classification on our test set.

Model	Accuracy	macro Precision	weighted Precision	macro Recall	weighted Recall	macro F1	weighted F1
RoBERTa	82.54%	83.32%	83.23%	80.03%	82.93%	81.02%	82.60%
BERT	73.87%	73.14%	74.85%	70.49%	75.61%	70.72%	74.43%
ALBERT	76.90%	77.79%	78.01%	71.91%	77.24%	72.51%	75.90%
DistilBERT	71.11%	72.36%	73.17%	65.81%	72.36%	64.98%	69.75%

[55], and RoBERTa-base [61] for this classification task. We found that RoBERTa-base outperformed the others on the test set regarding accuracy, precision, recall, and F1 score (Table 3). RoBERTa-base is a robustly optimized model using BERT-style pretraining methods and achieved the start-of-the-art results on downstream sequence classification tasks (e.g., 94.8% accuracy in the SST-2 sentiment analysis dataset) in 2019 [61]. We finetuned this model in our dataset using a batch size of 32 and 20 epochs, with an early stopping mechanism that selects the model with the highest accuracy on the validation set during the training process. The warm-up steps were 500, and the learning rate was 5e-5. We exploited a weight decay coefficient of 0.01 for regularization and the cross-entropy loss function. After fine-tuning, our sentence classifier achieves a 0.83 accuracy, a 0.83 weighted F1, and a 0.81 macro F1 score on the test set, indicating its appropriate validity. Table 4 shows the classification results on eight randomly sampled sentences. Our master student, who is experienced in UI design, suggested that sentences No.1-4 and No.6 were accurately classified, No.5 should be a rationale, No.7 should be a suggestion, and No.8 had better be treated as a critique. Our feedback sentence classifier has a chance to make an incorrect or incomplete decision. Nevertheless, it is overall acceptable in our usage scenario with the goal of helping users locate and digest meaningful feedback.

3.4 Keywords Recognition and Clustering

To help users explore comments based on their interests, our computational workflow further processes the sentences in the comment summary and outputs 1) the recognized keywords about UI or visual elements and 2) associated clusters that can reflect a higher-level design concept like

Table 4. Examples of the unlabeled sentences from the comment summary and corresponding classification results from our model.

No.	Sentence	Classification
1	With traditional news apps, users can scroll through content and find things that are interesting to them (or relevant to them).	rationale
2	One thing would be maybe adding a + icon to the 'add' button and shortening the width a bit?.	suggestion
3	The illustration in the last mock seems a little out of place considering the other mocks.	critique
4	My one suggestion would be to divide the card number, expiry and cvc into visually separate input boxes.	suggestion
5	This would give that big chunk of real-estate a purpose.	critique
6	The space can be used more effectively.	rationale
7	Title of the left section should match the tab/step you're in ("Payment Details")	critique
8	I don't see any point in including 4K labels on the movie unless you're also an app for booking virtual viewing.	suggestion

color and space [19]. After this step, each classified feedback sentence is tagged with the types of its UI components (Table 6) and visual elements mentioned (Table 7).

3.4.1 Keywords recognition. We approximate it as a token classification task following the related tutorial in Huggingface [30] which details how to fine-tune the pre-trained “bert-base-cased” language model [24] for name entity recognition. Specifically, we attributed labels “B-ui” / “B-visual” to the tokens at the beginning of relevant keywords, “I-ui” / “I-visual” to those inside the keywords, and “O” to the others. For this supervised learning task, the labeled dataset contains 828 sentences (i.e., those labeled critiques, suggestions, or rationales), which are tokenized and tagged with the BIO labels based on the 837 UI components and 583 visual element keywords. We used 680 sentences (80% of the data) for training and the rest sentences (20%) for testing. We evaluated the resulting model by considering both classified “B-ui” and “I-ui” tokens as UI components and both “B-visual” and “I-visual” tokens as visual elements, since these tokens, regardless of their positions in the keywords, can help users locate the content of their interests in the comments. It achieved a 0.88 accuracy, a 0.71 F1, a 0.80 precision, and a 0.64 recall score on the test set, suggesting an acceptable performance for a classification problem. Table 5 shows the recognition results of eight randomly selected sentences. Our master student, who is experienced in UI design, suggested that the UI components and visual elements in sentences No.1-5 and No.8 were correctly detected. However, the model made mistakes of classifying “the” as a UI component (which could be avoided with a rule to ignore words like “the” and “a”) in sentence No.6 and forgetting the UI component “labels” in sentence No.7. In general, our element keywords recognition model has a proper performance that can power our DesignQuizzer to personalize the learning materials.

3.4.2 Element keywords Clustering. We further attach the detected keywords with tags that can reflect a higher-level design concept, such as “color” and “space”, to support learning with a specific focus. We first adopted the pre-trained language model with the most downloads in the Huggingface’s Sentence Transformer tutorial, “all-MiniLM-L6-v2”, to map the keywords into a 384-dimensional dense vector space that can be used for tasks like clustering [28]. Then, we applied the well-developed K-means clustering algorithm to the semantic embeddings of UI components and visual elements, respectively. We manually went through the output results with the $K \in [3, 10]$ and found that most of the clusters make sense with $K = 4$ for grouping visual elements and with $K = 9$ for grouping UI components. Again, we invited our master student with UI design experience

Table 5. Examples of the recognized UI components and visual elements from sentences.

No.	Sentence	Output UI component	Output visual element
1	At first glance, i think you should make the call to action buy button a brighter color to catch peoples eyes.	call to action, buy button	brighter color
2	It doesn't have to be real website to be properly designed...	-	-
3	The heading again needs more contrast	heading	contrast
4	Gradient/contrast on the progress track (top of screen)looks weird, this should be more clear, lines are too thickContent columns/containers/boundaries unclear and unbalanced.	the progress track, lines	Gradient, contrast, boundaries
5	Things I'd change are, font weight in the CTA's plus the Add button in the investment block make it aligned with the rest of the text fields.	CTA's, Add button, investment block, text fields	-
6	It looks like the left column is slightly more narrow than the right.	the, left column, right	-
7	I remember seeing spelled-out "close" buttons on the left, but not as an icon.	"close" buttons, icon	-
8	I like the overall structure and layout.	-	structure, layout

Table 6. UI component clusters with representative keywords.

No	Clusters	Keywords
1	button-related	button, sign up button, login button, button background, close buttons, buy button, white button
2	text-related	lines, right-hand content, text, the second line of text, fonts, body text, content
3	card-related	badges, "Your Cart", card number, item card, price, shopping cart, product card
4	divider-related	labels, slider, dividers, the back arrow, the left of the image, dividing line, filter
5	color-related	colors, color palette, "add colour", "create colour", the color settings, background elements, dark theme
6	menus-related	title of the left section, menu options, clickable links, this page, drop-down menu, tab bar, header
7	icon-related	"trash bin icon", "lock icon", the ice cream icon, tab bar icons, iconography, profile icon, logos
8	general element	elements, container element, components, UI component, Control elements
9	others	the title of the post, username, planes, stats, ads, the progress track, delivery

Table 7. Visual element clusters with representative keywords.

No.	Cluster	Example keywords
1	space, shape	padding, space, whitespace, shape, rounded edges, align, margins, consistency, line height, width, floating
2	layout	layout, responsive layout, second screen, accessibility, information architecture, user flows, readability
3	typography	typography, fonts, the visual hierarchy of text, contrast, styling, the font size, visible, saturation, sans-serif
4	color	color, calm colors, lighter, black, color usage, medium gray, red, background color, yellow, dark, pink

to select one or two representative words that reflect the main higher-level design concepts of each cluster. Table 6 and Table 7 present the resulting clusters.

4 DESIGNQUIZZER SYSTEM

To facilitate users in visual design learning activities, we develop DesignQuizzer which prompts questions on the UI designs in online communities and presents structured feedback for explanations. We choose a quiz-based interaction design as previous learning support tools (e.g., QuizBot [82], Sara [98], CReBot [77]) have demonstrated its engagement and effectiveness. As our objective is to allow learning anywhere at any time, we build DesignQuizzer as a responsive web-based application (Figure 3). Further, to support learning at scale based on users' interests, DesignQuizzer should computationally generate a quiz pool that covers different types of visual elements. Below, we demonstrate the DesignQuizzer's design and evaluation with the processed data from Reddit r/UI_Design and the learners' tasks as getting familiar with "space", "shape" (cluster No.1 in Table 7), "typography" (No.3), "color" (No.4). We discuss the extended applications and limitations of the DesignQuizzer in section 7.

4.1 DesignQuizzer Interface and Interaction Design

As shown in Figure 3, the left-hand side is a chat window for users to interact with DesignQuizzer. To interact with the DesignQuizzer, users can click one of the buttons below its message or type down keywords if they want to explore a visual element or a UI component. The reasoning behind this mixed modality is to ensure both flexibility and efficiency regarding user interactions with the DesignQuizzer. Besides, these buttons help to maintain the conversational flow [11]. The right-hand side of the interface presents the title (part e) and the attached UI example (g) of the original feedback-request post to provide the main learning context. Users can click the UI example to view it in full screen and enlarge it to check its details. Part f offers a simple way for users to keep track of their quiz progress and performance [82, 93].

Following [82], Figure 4 illustrates one round of the interaction between a user and DesignQuizzer. At the beginning of each round, DesignQuizzer asks a user a single-choice question (part a in Figure 3) on the UI example selected according to our quiz strategy (subsection 4.3). The question masks the keywords about visual elements in a sentence that critiques the UI example (e.g., "the meaning for ___ and red for the pins is not clear"), and the options consist of the masked keywords (e.g., "grey") and another two words (e.g., "black" and "blues") as distractors. As our focus is to help users learn about visual elements, the critique with such a masked keyword used as a quiz prompt could direct their attention to a specific visual aspect of the UI example, and the options would encourage them to think of and differentiate related visual elements. We choose single-choice questions instead of open-ended questions to allow an easy way for user's input and ensure the system's robustness. A user can click one of the three options, click the "I don't know" button and confirm whether "give up" or not, or click the "I need a hint" button. If users ask for a hint, DesignQuizzer will give more contexts about the critique sentence by presenting the sentences before and behind it in the original comment (part b). The DesignQuizzer will provide positive reinforcement feedback that is typical of a study partner if a user chooses the correct answer, and it will give encouragement if a user chooses a incorrect option or gives up. After the user sees the correct answer (part c), they can click the "Why" button to check the structured summary of the comment that may provide suggestions and rationales for the critiques (d). To facilitate the visual search for users' interested content, we encode the critique, suggestion, rationale, UI component, and visual element in the comment summary with different colors. At the end of each quiz, the user can click the "I want to explore a UI component" or "I want to explore a visual element" button to type down and send related keywords. The DesignQuizzer will then retrieve a question about the

I got a critique of an example for you about "pins".

(a) "the meaning for ____ and red for the pins is not clear." Which of the following options would be the right one?

grey **black** **blues** **I don't know.** **I need a hint.**

I need a hint.

(b) You can check more information about this critique:
Needs to be more distinctive using unique elements not (just) change in colour (ex. "the meaning for ____ and red for the pins is not clear." Why the difference?)

grey **black** **blues** **I don't know.**

grey

Well done!

(c) The correct choice is: **grey**

Why? **Wait, it's not necessary right!** **Next Question!** **I want to explore an UI component.** **I want to explore a visual element.**

Why?

(d) Here is the related comment summary:

(critique) Otherwise **low contrast** on a lot of text, icons. **(critique)** Selected state for the **pin** on the map of screen 2 is not clear. **(suggestion)** Needs to be more distinctive using **unique elements** not (just) change in colour (ex. Larger w pronounced shadow or glowing edges etc.)" **(critique)** **the meaning for grey and red for the pins is not clear.**" **(suggestion)** I would repeat the **ice cream icon** in the **header** section in screen 2. Little reason to omit it then have it appear in screen **(suggestion)** Think about the animation when the **card** is pulled to full, you don't want the **text** to re-shuffle; enlarge text is fine as long as it stays at similar location in the **layout**, but avoid shuffling
For more about this UI example, please check this [link](#).

Next Question! **I want to explore an UI component.** **I want to explore a visual element.**

Select a button or type your message if needed... **Send**

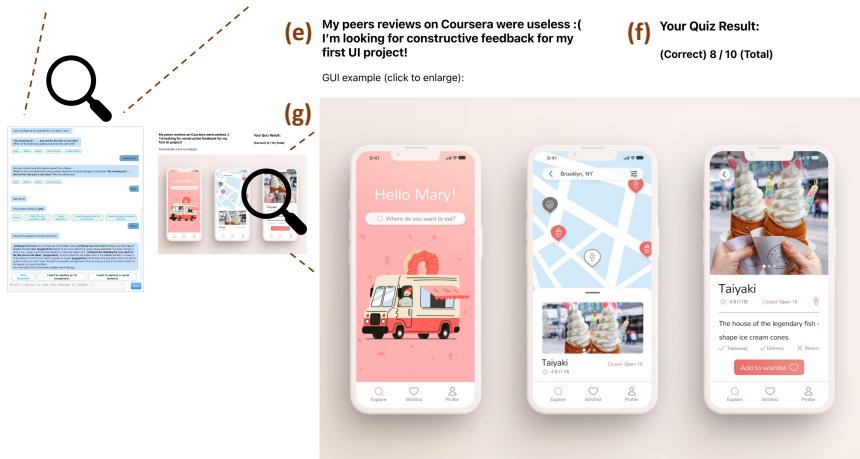


Fig. 3. Interface of DesignQuizzer with a sample UI design and a conversation log. (a) Question. (b) Hint. (c) Answer. (d) Structured comment summary. (e) Post title. (f) Performance track. (g) UI example.

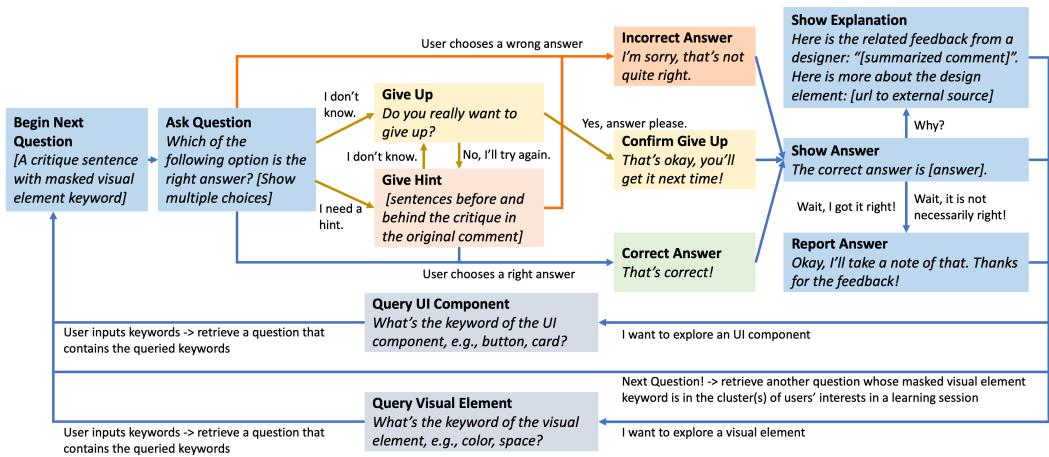


Fig. 4. The dialogue flow of DesignQuizzer including typical sample responses.

queried keywords and the associated UI example as the next quiz. To simulate the experience of human-human conversations, we provide a variety of different responses of DesignQuizzer at each state in the dialogue flow.

4.2 Quiz Pool

Our quiz pool contains single-choice questions about visual elements in UI design examples (Table 7). Every question has five parts: a feedback-request post with a title and a UI example, a critique that has a keyword of visual elements masked, options that contain the masked keyword and two distractors, a hint that includes the sentences around the critique, and a structured comment summary for providing explanations. Different from previous pedagogical conversational agents (e.g., [82, 98]) that need manual collections of the questions and their answers, we prepare the quiz pool by applying our trained computational models to the comments in online design communities. The procedure of the quiz preparation is detailed below:

1) Get feedback-request posts with downloadable UI examples. We first used the PushShift api to collect all the 1521 posts with the “Feedback Request” flair and with at least one comment (excluding the one from the bot AutoModerator and those from original posters) within 2019.5.30 and 2022.5.30 in Reddit r/UI_Design. Then we kept those posts attached with downloadable UI design images by checking if the post body contains an image link and if the image is publicly available. This step results in 502 posts, and each is associated with a UI design example. To reduce the human workload for the semi-automatic question and option preparation in step 3 below, we randomly select half of them, i.e., 251 posts, to continue with.

2) Process the comments with our computational models. Next, we applied our computational workflow (Figure 2) to all the comments of the sampled 251 posts. Note that the input to the feedback sentence classifier is a full sentence separated by “”, “?” or “!” in this case rather than the sub-sentences (e.g., can be part of a full sentence) that we labeled and were used to train the classifier. This is because we want to provide learners with a concrete context via full sentences in the questions and explanations. In total, there are 4437 comments on these 251 posts, which include 867 critiques, 2103 suggestions, 189 rationales sentences, 8171 occurrences of UI component keywords, and 5115 occurrences of visual element keywords. The unbalanced numbers between the suggestion and rationale sentences could be due to the fact that some full sentences may contain

both suggestion and rationale sub-sentences, while our model can not give multiple labels for one input sentence and favor labeling them as suggestions.

3) Prepare questions and options. We generated a multiple-choice question for each of the 867 critiques by masking the first keyword about visual elements if it exists, resulting in 509 questions. We adopted a semi-automatic method to prepare the distractors for these 509 questions. First, from the 5115 visual element keywords, we randomly selected two different words that fall into the same cluster (Table 7) and have the same POS tag (assigned by the nltk package) with the right answer for each question. Then, the first author manually went through the 509 questions and their options to validate the quality and make revisions. He identified 28 generated questions that do not make sense, e.g., the sentence is not a critique or the masked word is not about visual elements. Among the 962 distractors of the rest 481 questions, he revised 373 items that could be easily excluded by checking the grammar of the question.

4) Prepare hints and explanations. To prepare hints and explanations for users, we further processed the 481 questions and kept 337 of them that had at least one suggestion or one rationale sentence in the sourced comment of the question. **After the full procedure, we have 337 questions from the critiques of 152 posts.** Among them, the masked keywords of 133 questions are about visual elements “space” and “shape” (cluster No.1 in Table 7), 23 are about “layout”, 92 are about “typography”, and 89 items are about “color”. The numbers of unique keywords under each visual element cluster are 84, 13, 47, and 39, respectively.

4.3 Quiz Strategy and System Implementation

The DesignQuizzer’s strategy for selecting the next quiz depends on three conditions. First, if users click the “Next Question” button, it will randomly retrieve another question that aligns with users’ interests in a learning session. To be more specific, this question contains (masked) visual element keywords related to the target concept to learn but can be about any type (Table 6) of UI component in our dataset. For example, if a user wants to learn visual elements about “space” and “shape” in the learning session (one task in the later user study), DesignQuizzer’s next quiz will come from the 133 questions whose masked keywords are in cluster No.1 (Table 7). If a user wants to get familiar with “typography” and “color” (the other task), DesignQuizzer will retrieve a quiz from the 92 questions with answers in cluster No.3 or the 89 questions with answers in cluster No.4. Second, if users click the “I want to explore a visual element” button, DesignQuizzer will have a 50% chance to present a question with the queried keywords as the right answer and the other 50% chance to get a question whose answer is in the same visual element cluster of the queried word. This design choice could satisfy users’ interests while not making the quiz too easy for them. Third, if users click the “I want to explore a UI component” button, DesignQuizzer will retrieve a question that contains the queried keywords. For example, if a user queries “icons”, the Quizzer will randomly prompt a question that originates from a critique sentence containing icon-related UI component keywords (Table 6). In our quiz pool for user study, there are 30, 0, 19, and 16 icon-related questions that mention visual elements about space/shape, layout, typography, and color, respectively (Table 7). If there are no satisfactory questions in the second and third conditions, DesignQuizzer will apologize for not having related quizzes and switch to the first condition. To mitigate this issue, DesignQuizzer will randomly suggest two candidate keywords for references.

We implement DesignQuizzer’s frontend based on React.js and the react-viewer and chat-ui-kit-react components³. Its backend server is based on a python flask framework to maintain the

³Links of the main components: <https://github.com/infeng/react-viewer>, and <https://github.com/chatscope/chat-ui-kit-react>

dialogue flow. The UI design images are locally stored using python http.server package. We use the ElasticSearch engine to store the quiz pool and retrieve quizzes.

5 EXPERIMENT I

We conducted two experiments with novice designers to study DesignQuizzer’s benefits and drawbacks when compared with the community-like baseline interface for learning visual design. The experiment I with 24 participants adopts a within-subjects study design to reduce the possible effect of individual differences in the learning process and user perceptions of the tools. We counterbalance the order of used interfaces (DesignQuizzer vs. baseline) to mitigate the learning effects and confirm that the order does not significantly impact the reported results, as detailed in [subsubsection 5.5.1](#). Based on the insights from experiment I, experiment II with 28 participants further evaluates what users can learn with DesignQuizzer and how well they can apply the learned knowledge in a subsequent design activity. This experiment adopts a between-subjects design as the knowledge transition from one learning task to the other could largely affect the measured outcome of knowledge application. In this section, we present the design and results of experiment I. Our research questions are:

RQ1: How would DesignQuizzer affect the amount and the novices’ perceptions of their explored design examples and comments in the learning session?

RQ2: How would DesignQuizzer affect novices’ engagement and cognitive load in the learning session?

RQ3: How would DesignQuizzer affect the novices’ outcome on recognizing enacted or violated principles about the learned visual elements in others’ design examples?

RQ4: How would novices interact and perceive with DesignQuizzer in their learning sessions?

5.1 Experimental Design

Our experiment is a within-subject design. Each participant has one learning task with DesignQuizzer and the other with a baseline tool.

5.1.1 Baseline: Reddit r/UI_Design Interface. To simulate how users normally explore examples and comments in online design communities, we use the Reddit r/UI_Design webpage as the baseline ([Figure 1](#)). We use the Reddit r/UI_Design webpage as the baseline for two reasons. First, exploring examples and comments in online design communities is a common learning practice [4, 16, 32, 36, 107], and as stated in [subsubsection 3.1.1](#), r/UI_Design is a representative one for learning visual design. Second, since the current quiz pool of DesignQuizzer is sourced from the r/UI_Design community, using it as the baseline can ensure the coherence of learning materials. Participants can use its embedded functions to search the posts and comments inside the communities based on keywords and sort them based on “Hot”, “New”, “Top”, and “Rising”. We use the “Filter by Flair” feature to list all “Feedback Request” posts on the webpage in advance to avoid distraction from other types of posts, e.g., “Portfolio Review Requests”. Participants in the DesignQuizzer condition can also access this community webpage, e.g., via a link for more explanations ([Figure 3 part d](#)).

5.1.2 Learning tasks. Each participant has two learning tasks. The task prompt is: “You will learn design knowledge about (task A) ‘space’ and ‘shape’ / (task B) ‘color’ and ‘typography’ in this session. You will explore the online UI examples and comments to learn how these concepts are enacted or violated in practice. After the learning session, you need to critique a new UI example from the aspects of (task A) ‘space’ and ‘shape’ / (task B) ‘color’ and ‘typography’, give suggestions for improvement, and provide corresponding rationales if any”. Our quiz pool contains 84 unique keywords about “space” and “shape” and 86 unique keywords about “color” and “typography”,

which helps to balance the difficulty of the two tasks. DesignQuizzer will adjust the quiz strategy for “Next Question” (detailed in subsection 4.3) to support users’ learning goal in their task, while in the baseline condition users are presented with all feedback-request posts and comments in the Reddit interface. After being counterbalanced with Latin Square, there are four task-interface combinations, each with six participants: (a) task A (DesignQuizzer) - task B (Baseline), (b) task B (DesignQuizzer) - task A (Baseline), (c) task A (Baseline) - task B (DesignQuizzer) and (d) task B (Baseline) - task A (DesignQuizzer).

5.2 Participants

We recruited 24 students (10 Females, 11 Males, 3 Not Available; age range 18-21, $Mean = 19.25$, $SD = 0.74$; noted as P1-24) from a local university via a post in a group chat and word of mouth. The inclusion criteria are that participants are novices in visual design but have interests to learn more about it. We do not require the participants to be design students, because our Design-Tutor would support any students who are interested in learning visual design online. All of them are undergraduates and have passed the national College English Test for general requirements. Twenty major in Artificial Intelligence, two in Software Engineering, one in Micro-electronics, and one in Ocean Engineering. In general, our participants have little or no experience in learning UI visual design ($M = 1.42$, $SD = 0.88$) and exploring online design communities ($M = 2.83$, $SD = 1.90$; 1 - No experience at all, 7 - A lot of experience) but they are interested in learning it in our study ($M = 5.42$, $SD = 1.18$; 1 - No interest at all, 7 - A great deal of interest). The usage frequency of online communities where people create posts and comments is: 17 daily, 5 2-6 days a week, 1 once a week, and 1 less than once a week.

5.3 Measures

RQ1. Explored design examples and comments. We log the numbers of explored design examples and comments by analyzing the interaction log of DesignQuizzer and the screen video record of the Reddit condition, e.g., plus one to the numbers if users’ web page stays in the unique example or comment for more than four seconds as indicated by the authors’ trials. Besides, after each task, we measure participants’ satisfaction and perceived helpfulness of the explored examples and comments using two 7-point Likert scale items (1 - strongly disagree; 7 - strongly agree) adapted from [75, 76].

RQ2. Engagement in the learning process. We derive six 7-point Likert scale items (Cronbach’s $\alpha = 0.793$) to measure user engagement based on Brien’s theoretical model [72] regarding the flow theory for a positive experience [21]. Specifically, they are concentration (“completely involved, focused, and concentrating”), a sense of ecstasy (“feel doing something is special”), doability (“skills are adequate, neither anxious nor bored”), sense of serenity (“forgot about myself doing something”), timeless feeling (“time passed quickly”), and intrinsic motivation (“feel self-rewarded”) ⁴. As for the cognitive load, we use one 7-point item “I think the cognitive load of exploring examples and comments to learn the targeted visual design elements with this tool is very high” (1 - strongly disagree, 7 - strongly agree) adapted from [41].

RQ3. Outcome of the learning session on recognizing enacted or violated design principles. Because online design communities like Reddit do not make any systematic attempt to measure learning, a user study like ours must rely heavily on proxy measures of the learning outcomes. Our measure for the development of visual design skills is inspired by previous quantitative studies on computational learning in the Scratch community, which capture learning gain via the size of the learners’ repertoire in terms of the number of different types of computational

⁴The full items are provided in the supplementary materials.

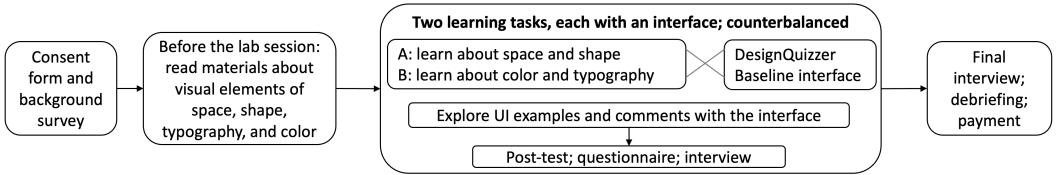


Fig. 5. Procedure of the within-subjects (tool: DesignQuizzer vs. Baseline Reddit interface) experiment I. In each learning task, participants explore design examples and comments with the assigned interface to learn the required visual elements.

thinking concepts a user has demonstrated in their projects [14, 22, 23, 84, 106]. Specifically, we construct an outcome measure on the numbers of enacted or violated principles about the visual elements (i.e., space and layout in one condition, color and typography in the other) mentioned in the participants' feedback comments to the unseen UI examples in each post-test. For instance, "LOGO and navigation bar are located on the same line and aligned" in the comment counts as one point as it recognizes the enacted layout alignment principles of the design example. Two annotators first separately labeled 12 randomly sampled comments and then discussed and reached a consensus on the rating scheme, e.g., the sentence should include a specific UI component or visual element. Next, they independently labeled all 48 comments in the post-test; ICC(intraclass correlation coefficient) = 0.856, $p < 0.001$. They resolved the disagreement by discussions.

RQ4. Interaction and perception with DesignQuizzer. We log the number of participants' clicks on each button to understand how participants use DesignQuizzer. We adapt the technology acceptance model [92] that has been used for evaluating educational chatbots [77, 93] to measure the following in each condition: usefulness (four items, Cronbach's $\alpha = 0.897$); easy to use (four items, Cronbach's $\alpha = 0.737$); and intention to use (two items, Cronbach's $\alpha = 0.782$). We average the ratings of multiple questions as the final score for each factor in the acceptance model.

5.4 Procedure

Figure 5 illustrates the procedure of experiment I. After obtaining participants' consent, we sent them the links to an online background survey and a document that lists the basic definitions and concepts of visual elements about space, shape, typography, and color. The document contains 13 screenshots from related pages of [37] and [19] that introduce the basic design terminology. At the beginning of each task, we first introduced the task and the tool interface. Then participants explored online UI examples and comments with the assigned tool. We allocated 25 minutes for each learning session based on a pilot study with two users. After each task, participants wrote a comment to criticize a new UI example in the post-test and filled in the questionnaire to rate their perceptions of the learning process and the tool. We further conducted a semi-structured interview to make sense of the ratings and suggestions to improve the tool. Upon completion of two tasks, we asked which tool they preferred and why. The whole procedure lasted for 100-120 minutes. After debriefing, each participant received a \$8.5 compensation following the local payment policy.

5.5 Analysis and Results

For the self-report items and numbers of enacted or violated principles about the visual elements mentioned in the post-tests' comments, we performed Wilcoxon signed-rank tests [100], as used in previous HCI studies [46, 97, 103], to assess the difference between the DesignQuizzer and baseline conditions (Table 8). We also conducted a set of statistical tests (detailed in subsubsection 5.5.1) to affirm that neither the factor of tool/task orders nor their interactions with the tool factor impact

Table 8. The experiment I's RQ1-4 statistical results about DesignQuizzer and the community-like baseline interface. All items except the one for RQ3 are measured using a standard 7-point Likert scale (1 - strongly disagree; 7 - strongly agree). Note: -: $p > .1$, + : $.05 < p < .10$, * : $p < .05$, ** : $p < .01$, *** : $p < .001$; Wilcoxon signed-rank test; within-subjects; $N = 24$.

Research Question	Item	DesignQuizzer	Baseline	Statistics		
		Mean (SD)	Mean (SD)	Z	p	Sig.
(RQ1) Explored examples and comments	Satisfaction	5.29 (0.86)	4.67 (1.61)	-1.774	0.076	+
	Helpfulness	5.83 (1.20)	3.83 (1.61)	-3.453	0.001	**
(RQ2) Engagement and cognitive load in the process	Mean engagement	5.35 (0.82)	4.39 (1.14)	-3.137	0.002	**
	- Concentration	5.67 (1.24)	4.24 (1.51)	-2.890	0.004	**
	- Sense of Ecstasy	6.25 (0.85)	4.33 (1.52)	-4.073	0.000	***
	- Doability	5.04 (1.37)	4.75 (1.42)	-0.821	0.412	-
	- Sense of Serenity	4.38 (1.61)	4.25 (1.94)	-0.292	0.771	-
	- Timelessness Feeling	5.50 (1.29)	4.50 (1.79)	-2.027	0.043	*
	- Intrinsic Motivation	5.29 (1.30)	4.25 (1.54)	-2.427	0.015	*
(RQ3) Outcome on recognizing enacted or violated design principles	Cognitive load	5.04 (1.04)	4.42 (1.25)	-1.557	0.119	-
	Numbers of design principles about visual elements mentioned in the post-test's comment	3.38 (1.50)	2.29 (1.23)	-2.994	0.003	**
(RQ4) Perception towards the tool	Usefulness	5.65 (0.80)	4.38 (1.08)	-3.754	0.000	***
	Easy to use	5.29 (0.97)	4.13 (1.10)	-3.411	0.001	**
	Intention to use	5.46 (1.01)	4.54 (1.06)	-2.725	0.006	**

Table 9. Summarized pros and cons of DesignQuizzer and the community-like baseline interface in experiment I. These findings are incorporated into subsubsections 5.5.2 - 5.5.5 to make sense of the statistical results. The number next to each point is the number of participants who mention it; within-subjects; $N = 24$.

	DesignQuizzer	Baseline
Pros	Related comments (6); "Query" function (11); Quiz-like interaction (16); Facilitate deep understanding (4); Structured comment summary (11); Clear interaction (7)	Sense of a community (9); Entertaining content (5); Broad knowledge view (3)
Cons	Boring interaction (3); Lack external knowledge (4); Insufficient explanation in the summary (7); Overwhelming visual encodings (5)	Unrelated comments (15); Low-quality comments (5); Passive learning experience (7); Unstructured comments (5); Lack external knowledge (4); Unaligned comments to the UI image (4); Messing comment structure (4)

the significances of the reported results. This indicates that the learning bias did not significantly impact our results. For the interview recordings, two of the authors transcribed them into text. They first familiarized themselves by reviewing all the text scripts independently. After several rounds of coding with comparison and discussion, they finalized the codes of all the interview data regarding each RQ aspect. We counted the occurrences of codes (Table 9) and incorporated these qualitative findings in the following presentation of our results.

5.5.1 The Impact of Tool/Task Orders on the Measures. We first conducted a set of mixed ANOVA analyses using SPSS software to examine the impact of tool/task order on our measures. The within-subjects factor is the used tool (DesignQuizzer vs. baseline interface), and the between-subjects factors are the order of the learning task and the order of the experienced tool (Figure 5). As shown

Table 10. The p-values of the mixed-ANOVA tests to examine the impact of tool/task orders on the measures in experiment I. Within-subjects: **tool** (DesignQuizzer vs. baseline); between-subjects: the **order** of the used tool, the order of learning **task**; $N = 24$. The tests affirm that the orders do not significantly impact the significance of the reported results.

	Within-subjects effects				Between-subjects effects		
	Tool	Tool * Task	Tool * Order	Tool * Task * Order	Task	Order	Task * Order
Satisfaction	0.075	0.538	0.538	0.711	0.017	0.319	0.155
Helpfulness	0.000	0.844	0.844	1.000	0.194	0.850	0.451
Mean engagement	0.001	0.597	0.889	0.090	0.304	0.248	0.137
Cognitive load	0.107	0.739	0.739	0.580	0.897	0.165	0.698
Numbers of design principles	0.002	0.793	0.600	1.000	0.325	0.619	0.868
Usefulness	0.000	0.538	0.194	0.538	0.119	0.193	0.301
Easy to use	0.000	0.473	0.139	0.631	0.173	0.586	0.903
Intention to use	0.003	0.112	0.085	0.665	0.519	0.605	0.519

in Table 10, most of the main effects of tool/task orders and their interaction effects with the tool condition are not significant. One exception is the effect ($F = 6.835, p = 0.017$) of the task order factor on perceived satisfaction. Specifically, the twelve participants with the task order “B (color and typography) -> A (space and shape)” ($M = 5.46$) were generally more satisfied with the explored design examples and comments compared to the other twelve participants with “A -> B” task order ($M = 4.50$). Nevertheless, the ordering effects do not affect our key findings on helpfulness, engagement, number of design principles, and perceived usefulness. Therefore, in the following reported results, we focus on the effects of the within-subjects tool factor. While ANOVA helps mixed analyses for testing the ordering effect, it is generally used for the comparisons of more than two means [38]. In our case, we performed Wilcoxon signed-rank tests [100] to assess the difference between the DesignQuizzer and the community-like baseline conditions.

5.5.2 Explored Design Examples and Comments (RQ1). In general, participants explored more UI examples ($Mean = 13.21, SD = 6.62$ vs. $M = 6.71, SD = 1.55$) and comments ($M = 20.67, SD = 10.07$ vs. $M = 13.71, SD = 5.53$) with DesignQuizzer than with the baseline tool during the 25-minutes learning session. Overall, there is a tendency that participants are more satisfied with the explored examples and comments for learning required visual elements ($Median = 5.00$) than in the baseline condition ($Mdn = 5.00$); $Z = -1.774, p = 0.076$. Compared with the baseline condition ($Mdn = 4.00$), they felt that most of their explored examples and comments were significantly more helpful for their learning goals in the DesignQuizzer condition ($Mdn = 6.00$); $Z = -3.453, p = 0.001$. Fifteen users mentioned that most of the Reddit comments were unrelated to their interested visual elements. Five of them reported that they encountered low-quality comments in Reddit. “*It is hard to find the needed information, especially from long comments, in the Reddit community*” (P19, Female, age: 19). Six people reported that DesignQuizzer’s prompted examples and comments matched their learning interests, and eleven people appreciated its “query” function. “*I like the DesignQuizzer’s ‘I want to explore a UI component / visual element’ buttons that allow me quickly explore examples and comments of interests*” (P7, Not Available, 20). These results suggest that DesignQuizzer helped users explore helpful UI design examples and comments more efficiently in comparison to the community-like interface.

5.5.3 Engagement in the Process (RQ2). In general, participants felt that they were significantly more engaged in the learning session with the DesignQuizzer condition ($Mdn = 5.50$) than with the baseline interface ($Mdn = 4.58$); $Z = -3.137, p = 0.002$ (Table 8). Specifically, DesignQuizzer

improves users' concentration ($Z = -2.890, p = 0.004$), sense of ecstasy ($Z = -4.073, p < 0.001$), timelessness feeling ($Z = -2.027, p = 0.043$), and intrinsic motivation ($Z = -2.427, p = 0.015$) during the visual design learning process. Sixteen participants highlighted that the quiz-like interaction of DesignQuizzer encouraged active thinking, which helped them focus on the learning materials. *"These questions made me more focused. I want to answer them correctly"* (P16, F, 19). *"DesignQuizzer's hint guided me to think actively by examining the image with the critique"* (P17, Male, 19). As for the Reddit interface, seven users reported having a passive learning experience, and five people encountered unstructured comments with scattered knowledge points. *"I gradually lost interest in learning because the resources in the comments are not organized and I was passively receiving them"* (P2, M, 21). Nevertheless, there is no significant difference between the tool conditions regarding the doability and sense of serenity. Participants' comments on the cons of DesignQuizzer and pros of Reddit baseline could explain this result. For example, three participants commented that they got bored with DesignQuizzer after having several rounds of interactions. *"I was tired of many single-choice questions, which are monotonous"* (P9, M, 19). In contrast, nine people mentioned that the Reddit design community offered them a *"relaxing learning environment with the sense of a community"*. Besides, five users liked the *"entertaining content"* (P15, M, 19) in Reddit, while DesignQuizzer filters out this content in the comment summarization step.

5.5.4 Outcome on recognizing enacted or violated design principles (RQ3). Compared to the conditions with the community-like interface ($Mdn = 3.00$), when using DesignQuizzer, participants mentioned significantly more points of design principles about the learned visual elements in their written comments to new examples in the post-test; $Z = -2.994, p = 0.003$ (Table 8). This suggests that compared to the baseline interface, DesignQuizzer could better help users develop their design skills in recognizing enacted or violated design principles about learned visual design elements in the UI examples. In the interview after each task, we asked participants if they could give examples of what they had learned in the learning session. Seventeen / twelve people provided such examples when they learned with the DesignQuizzer / Reddit interface. *"(With DesignQuizzer) I learned how to choose the shape of icons. Beginners of design may prefer to use pointed edges for the icons, but customers would prefer the rounded edges"* (P7, NA, 20). *"(With Reddit) I learned that we could use the hero title with a large font size to draw people's attention"* (P13, NA, 19). Four users mentioned that DesignQuizzer helped them gain a deep understanding of these elements, while three people indicated that the comment list of a UI design in the community provides a broad knowledge view. *"The DesignQuizzer helped me understand the design principles about color more deeply, especially when I answered the question incorrectly"* (P8, M, 18). *"The Reddit interface gave me a sense of learning in a community where the members' opinions are diverse and comprehensive"* (P11, M, 18). Nevertheless, the learning activity in both conditions could suffer from the lack of knowledge input from external resources, as commented by four participants. *"Some comments did not give reasonable rationales for the critiques. It would be better to provide the definition and a knowledge graph for the mentioned visual elements"* (P11, M, 18).

5.5.5 Interaction and Perception with DesignQuizzer (RQ4). Overall, participants actively responded to DesignQuizzer's questions ($M = 20.67, SD = 10.07$) and answer correctly for $13.54 (SD = 7.37)$ times. They seldom clicked the "I need a hint" button ($M = 2.13, SD = 2.38$) but frequently hit the "Why" button to assess the structured comment summary after seeing the correct answer ($M = 9.42, SD = 5.19$). Besides, they sometimes input their interested keywords via the "I want to explore a UI component / visual element" buttons ($M = 3.00, SD = 1.31$). In total, they reported the correctness of the answers eight times.

In terms of the perceptions of the tools, participants rate DesignQuizzer ($Mdn = 5.75$) to be significantly more useful for visual design learning than the community-like baseline ($Mdn = 4.50$);

$Z = -3.754, p < .001$ (**Table 8**). Eleven participants valued DesignQuizzer's structured comment summary. “*My favorite feature of DesignQuizzer is the ‘Why’ button, which provided me meaningful feedback sentences marked in different colors. I can also go to the original post thread with the provided link*” (P24, F, 19). However, seven users mentioned that the comment summary sometimes could not address their confusions, as P24 (F, 19) further said: “*But sometimes the comment summary did not provide sufficient information for me to understand why it is this answer. It would be better to have more explanations*”. Moreover, participants felt that DesignQuizzer ($Mdn = 5.50$) is significantly easier to use than the community-like interface ($Mdn = 4.13$) for visual design learning; $Z = -3.411, p = .001$. Seven people commented that the interaction with DesignQuizzer was “*clear, funny, and flexible*” (P1, F, 19). While the structured comment summary was valuable for eleven users, five suggested that its visual design could be further improved. “*The visual codes of the ‘why’ message were too much. Sometimes it was hard to find the key points as I did not know the meanings of these text colors*” (P18, M, 21). As for the Reddit community interface, four participants indicated that its layout design was not convenient for aligning the comments to the UI image, and four users felt that the nesting structure of comments was a mess for their learning purpose. “*The design example and comments are far from each other. I can not check them simultaneously but need to scroll the page up and down frequently*” (P7, NA, 20). All in all, participants have a significantly stronger intention to use DesignQuizzer ($Mdn = 5.50$) than the baseline tool ($Mdn = 4.50$) in their future informal learning practices of visual design; $Z = -2.725, p = .006$. In the interview after two tasks, all participants indicated their preferences for DesignQuizzer over the community-like interface for learning visual design. In addition to the benefits mentioned above, six people favored DesignQuizzer’s single-choice questioning design, which “*is easy to get started and reduces the learning cost of the tool*” (P9, M, 19). However, twelve participants were reminded that the preferences of DesignQuizzer depend on their roles as novices and their goals to learn specific visual elements. “*The Reddit webpage may be better if I want to broadly explore the examples and comments without a specific learning focus*” (P13, NA, 19).

5.5.6 Summary of the Findings in Experiment I. Compared with the community-like baseline interface, DesignQuizzer improves novices’ efficiency and overall engagement in exploring helpful UI examples and comments for learning specific visual elements. They favor its question-answering interaction that stimulates active thinking and its structured comment summary that reduces the reading workload. Besides, we find a significant improvement with DesignQuizzer in participants’ ability to critique others’ designs using the learned knowledge about targeted visual elements. However, our experiment I does not evaluate the impact of DesignQuizzer on participants’ ability to apply what they learn in future design activities, which is also an important aspect of learning design knowledge. Studying this aspect would require a between-subjects experiment design that is different from experiment I, because the knowledge transition from one learning task to the other could largely affect the participants’ performance in the subsequent design activities. Furthermore, we incorporate feedback from design experts on the proposed DesignQuizzer in the following experiment II to provide a more comprehensive assessment of the tool.

6 EXPERIMENT II

The primary goal of our between-subjects Experiment II with 28 participants is to explore an additional research question:

RQ5: How would DesignQuizzer affect the novices’ learning outcome on **a)** design knowledge and **b)** application of the knowledge in the visual design activity?

Specifically, we measure learning outcome on RQ5b via participants’ performance in the visual design tasks. This measure can reveal people’s skill in creating new or original work using the

learned knowledge, which is the most complex stage of the learning process compared to the “remember”, “understand”, “apply”, “analyse”, and “evaluate” stages in Bloom Taxonomy [65].

To complement the evaluation of DesignQuizzer, experiment II also serves the goals of inputting more evidence on the RQ1-4 results and collecting qualitative feedback from visual design experts.

The **baseline** Reddit r/UI_Design interface is identical to the one used in experiment I, as described in [subsubsection 5.1.1](#).

The **learning task** is similar to the one in experiment I, as described in [subsubsection 5.1.2](#). However, this time, each participant only needs to conduct one learning task: learn design knowledge about ‘color’ and ‘typography’.

We recruited 28 student **participants** (14 Females, 14 Males; age range 19-23, $Mean = 21.25$, $SD = 1.08$) via a recruitment post in a local university. We conducted a power analysis using the G*power software to determine the number of participants. Specifically, in the 2×2 mixed-design ANOVA test for the RQ5b’s measure described in [subsection 6.3](#), we input to G*power an effect size of 0.5 (expected means difference of measures: 1, $SD: 1$), a power of 0.8, and a p-value of 0.05. This outputs that the recommended smallest sample size is 26 (13 in each group). The inclusion criteria are that participants are novices in visual design but are interested in learning more about it. All of them have passed the National College English Test for general requirements. In general, our participants have little or no experience in learning UI visual design ($M = 2.00$, $SD = 1.09$) and exploring online design communities ($M = 3.18$, $SD = 1.85$; 1 - No experience at all, 7 - A lot of experience) but they are interested in learning it in our study ($M = 5.36$, $SD = 0.83$; 1 - No interest at all, 7 - A great deal of interest). The usage frequency of online communities where people create posts and comments is: 22 daily, 1 2-6 days a week, 3 once a week, and 2 less than once a week. We randomly assign our participants into the DesignQuizzer (noted as PD1-14) and Baseline groups (PB1-14), each with 14 participants.

6.1 Measures

We input more evidences on the RQ1-4 results regarding the following measures (detailed in [subsection 5.3](#)):

- **RQ1:** Satisfaction and perceived helpfulness of the explored examples and comments for learning design knowledge about color and typography.
- **RQ2:** Perceived engagement in the learning process.
- **RQ3:** Numbers of enacted or violated principles about color and typography mentioned in the participants’ feedback comments to others’ UI examples after learning session.
- **RQ4:** Perceived usefulness, easy to use, and intention to use regarding the DesignQuizzer and baseline interface.

We also offer more qualitative findings regarding **RQ4** about perception with the used tool. In the questionnaire after the learning task, we ask participants to write down their perceived pros/cons of the tool, the possible scenarios in their long-term usage of the tool, and suggestions for improvement.

To address the added **RQ5a** about the learning outcome on design knowledge, we ask participants in the questionnaire to list the knowledge related to color and typography learned from the explored UI examples and comments.

As for the **RQ5b** about applying learned knowledge in the design activity, we capture the change in participants’ performance on a visual design activity before and after the learning task. Specifically, in the **pre-test** before the learning task, participants are required to color the UI components and adjust the typography of two given uncolored mockups ([Figure 8a](#) and [Figure 9a](#)). This is a common graphic or visual design activity in design firms, in which designers usually deal

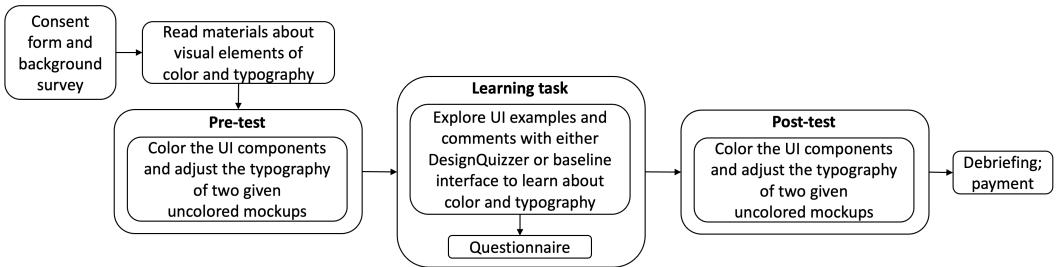


Fig. 6. Procedure of the between-subjects (tool: DesignQuizzer vs. Baseline Reddit interface) experiment II.

with a high-fidelity representation of a UI that shows exactly what the UI is supposed to look like (i.e., mockup) [69]. The goal of visual design is to literally enable proper visual communication of the UI's information using elements such as color, images, typography, and layout [69]. Therefore, this design activity enables us to capture the participants' performance in applying the knowledge about color and typography that they need to learn. We collect two mockups that have different types of UI components from the Figma design community⁵. One is a web page of a travel agency and has UI components like menu, button, and hero title. The other is a mobile payment page of an educational app and contains UI components like card, button, and form. We adjust the mockups as follows. First, we only keep one page of each original mockup to ease the design activity for our novice participants. Second, we uncolor the kept pages by assigning black, white, and grey colors to all UI components except the images (e.g., the background of the web page mockup). Third, we normalize the font style, size (=20), weight (regular) and color (# 000000) of the texts in each mockup.

In the **post-test** after the learning task, participants are required to conduct the same visual design activity on the same two uncolored mockups as those in the pre-test. We choose the same two mockups in pre- and post-tests to avoid noises from the design materials on the measured performance. We invite two visual design experts (E1-2, both males, ages: 28 and 27) to score the participants' outcome UIs from the visual design activity. E1 has experience in UI research projects and has been in the UI design teams of a travel agency company and a learning app startup. E2 majors in engineering design and has contributed to more than three UI design projects, including websites and corresponding mobile apps. They score each outcome UI from three aspects based on the suggested design principles from MaterialDesign [37]: **Consistent**: color and typography should be applied throughout a UI consistently and be compatible with the brand. **Distinct**: color and typography should create a distinction between elements, with sufficient contrast between them. **Intentional**: color and typography should be applied purposefully as it can convey meaning in multiple ways. We randomize the order of outcome UIs for the scoring session. The experts do not know whether the scoring UI is from the pre- or post-test and whether it is from the DesignQuizzer or Baseline group. For each aspect of an outcome UI, we average the two experts' scores (range: 0-10 points) as the final score.

6.2 Procedure

Figure 6 illustrates the procedure of experiment II. Participants used the desktop computer with a 23.8-inch monitor in our lab to conduct the study. After filling in the consent form and background survey, participants walked through the basic definitions and concepts of color and typography in

⁵<https://www.figma.com/community/file/1074977530633823056> and <https://www.figma.com/community/file/1079825061633332655>

a document compiled from [37] and [19]. We then demonstrated how to use the Figma desktop application to color the UI components and adjust the typography of a UI example. Next, participants conducted the pre-test in the visual design activity on two mockups. We allocated twelve minutes for the pre-test based on a pilot study with two users. After the pre-test, we introduced the used tool, i.e., either DesignQuizzer or Baseline, for the learning task. Participants started to explore online UI examples and comments with the assigned tool for 25 minutes. Upon completion of the learning task, participants filled in a questionnaire that asked them to 1) rate their perceptions of the learning process and the tool, 2) comment on a given new UI from the aspects of color and typography, and 3) write down what knowledge they learned, perceived pros/cons of the tool, how would they use it in the long term, and their suggestions for improvement. Finally, participants conducted the post-test in the visual design activity on the same two mockups as in pre-test within twelve minutes. The whole procedure lasted for 90-120 minutes. After debriefing, each participant received a \$12.5 compensation.

6.3 Analysis and Results

For the self-report items and numbers of enacted or violated principles about the visual elements mentioned in the post-tests' comments, we used the Mann-Whitney U test [66] to compare the ratings between two user groups (**Table 11**). The Mann-Whitney U is a non-parametric test commonly used to compare differences between independent conditions (e.g., in HCI studies [10, 48, 91]) especially when the data normality is violated, as confirmed in our cases. For the written responses in the questionnaire about RQ4 and RQ5a, two of the authors conducted several rounds of open coding with comparison and discussion. We counted the occurrences of codes and reported them in **Table 12** and **Table 13**. To evaluate the changes in the participants' performance in the visual design activity (**RQ5b**), we conducted a two-way mixed ANOVA to compare the performance of participants in each group (as the between-subjects factor) in the pre-test and post-test (as the within-subjects factor).

6.3.1 Statistical Results of RQ1-4. **Table 11** summarizes the statistical results of RQ1-4 in experiment II. Overall, there is no significant difference between the DesignQuizzer ($Mean = 5.07, SD = 1.07$) and baseline conditions ($M = 5.00, SD = 0.78$) regarding perceived satisfaction with explored examples and comments for learning color and typography; $U = 87.00, p = 0.592$. The perceived helpfulness of the explored examples and comments tends to be higher in DesignQuizzer ($M = 5.50, SD = 1.02$) than that in the baseline condition ($M = 4.79, SD = 1.42$); $U = 75.00, p = 0.272$. These results moderately support RQ1 findings in experiment I that DesignQuizzer could help users explore helpful UI design examples and comments for their learning tasks in comparison to the community-like interface.

In general, participants with DesignQuizzer ($M = 5.40, SD = 1.08$) in experiment II felt that they were significantly more engaged in the learning session than those with the baseline interface ($M = 4.67, SD = 0.80$); $U = 51.00, p = 0.030$. Specifically, DesignQuizzer significantly improved users' sense of ecstasy ($U = 34.00, p = 0.002$) and serenity ($U = 47.00, p = 0.017$) during the visual design learning process. There are no significant differences regarding other aspects of engagement and cognitive load between the two user groups. These results also moderately support RQ2 results in experiment I that compared to the community-like interface, DesignQuizzer can provide more engaging learning experience to novices.

On average, compared to the user group with the baseline interface ($M = 4.14, SD = 2.32$), the group with DesignQuizzer ($M = 3.29, SD = 1.44$) mentions more points of color and typography design principles in their comments to the new UI example. However, there is no significant difference between the two conditions regarding the learning outcome on recognizing enacted or

Table 11. The experiment II's RQ1-4 statistical results about DesignQuizzer and the community-like baseline interface. All items except the one for RQ3 are measured using a standard 7-point Likert scale (1 - strongly disagree; 7 - strongly agree). Note: -: $p > .1$, + : $.05 < p < .10$, * : $p < .05$, ** : $p < .01$; Mann-Whitney U test; between-subjects; $N = 28$.

Research Question	Item	DesignQuizzer	Baseline	Statistics		
		Mean (SD)	Mean (SD)	U	p	Sig.
(RQ1) Explored examples and comments	Satisfaction	5.07 (1.07)	5.00 (0.78)	87.00	0.592	-
	Helpfulness	5.50 (1.02)	4.79 (1.42)	75.00	0.272	-
(RQ2) Engagement and cognitive load in the process	Mean engagement	5.40 (1.08)	4.67 (0.80)	51.00	0.030	*
	- Concentration	5.71 (1.33)	5.43 (1.22)	83.00	0.469	-
	- Sense of Ecstasy	6.29 (0.61)	4.86 (1.41)	34.00	0.002	**
	- Doability	4.79 (1.25)	5.14 (0.86)	80.50	0.395	-
	- Sense of Serenity	4.93 (1.44)	3.47 (1.28)	47.00	0.017	*
	- Timelessness Feeling	5.21 (1.53)	4.57 (1.50)	62.50	0.097	+
	- Intrinsic Motivation	5.40 (1.08)	4.68 (0.80)	72.50	0.232	-
(RQ3) Outcome on recognizing enacted or violated design principles	Cognitive load	4.64 (1.22)	4.00 (1.62)	71.50	0.212	-
	Numbers of design principles about visual elements mentioned in the post-test's comment	4.14 (2.32)	3.29 (1.44)	91.50	0.756	-
(RQ4) Perception towards the tool	Usefulness	5.61 (0.94)	4.38 (0.96)	34.00	0.003	**
	Easy to use	5.13 (1.40)	3.95 (0.92)	40.50	0.008	**
	Intention to use	5.18 (1.27)	4.54 (1.63)	75.50	0.297	-

violated design principles; $U = 91.50$, $p = 0.756$. It moderately supports the RQ3 result in experiment I that DesignQuizzer could better help users develop their visual design skills in recognizing enacted or violated design principles about targeted visual elements.

Participants with DesignQuizzer ($M = 5.61$, $SD = 0.94$) rated it to be significantly more useful for visual design learning than those with the community-line baseline interface ($M = 4.38$, $SD = 0.96$); $U = 34.00$, $p = 0.003$. Besides, the DesignQuizzer users ($M = 5.13$, $SD = 1.40$) found it to be significantly easier to use for visual design learning than the baseline users ($M = 3.95$, $SD = 0.92$); $U = 40.50$, $p = 0.008$. On average, participants with DesignQuizzer ($M = 5.18$, $SD = 1.27$) had a higher intention to use it for future visual design learning than those with the baseline interface ($M = 4.54$, $SD = 1.63$); $U = 75.50$, $p = 0.297$. These results offer strong evidences to RQ4 results in experiment I that DesignQuizzer could be more useful and easier to use than the community-like interface for learning visual design knowledge.

6.3.2 Qualitative Results of RQ4. Table 12 offers more qualitative findings about users' perceived pros/cons of DesignQuizzer and baseline interface, the possible long-term usage scenarios, and suggestions for improvements.

Pros and cons. Similar to the qualitative feedback in experiment I (Table 9), participants with DesignQuizzer reported that the encountered comments are related to knowledge about color and typography (Number of participants who mention it = 5). They favored its quiz-like interaction ($N = 5$) and structured comment summary ($N = 3$) and felt that it facilitated a deep understanding of the knowledge points ($N = 4$). Four participants further commented that the Quizzer can help them focus the critical opinions on the UI design example. “*With DesignQuizzer, I can pay attention to the good and bad points of the design example and the specific reasons. It deepened my understanding and memory of the design details*” (PD8, Female, age: 20). However, similar to the findings in experiment I, some participants felt bored after several rounds of interaction with DesignQuizzer ($N = 3$), felt

Table 12. Summarized pros and cons of DesignQuizzer and the community-like baseline interface, possible long-term usage scenarios, and suggestions for improvement in experiment II. The number next to each point is the number of participants who mention it; between-subjects; $N = 28$.

	DesignQuizzer	Baseline
Pros	Related comments (5); Quiz-like interaction (5); Facilitate deep understanding (4); Structured comment summary (3); Focus on critiques (4)	Sense of a community (6); Entertaining content (1); Broad knowledge view (6); Diverse examples (2)
Cons	Boring interaction (3); Lack external knowledge (3); Insufficient explanation in the summary (6); Unable to select interested examples (1)	Unrelated comments (6); Low-quality comments (1); Conflicting opinions (1) Unaligned comments to the UI image (5); Messing comment structure (1)
Long-term usage	Poster, ppt, painting, photoshop (1); UI design project (7); Seek inspiration (2); Study at leisure time (3)	UI design project (3); Seek inspiration (3); Study at leisure time (3); Seek feedback (4)
Suggestion	Give feedback to users' designs (2); Involve structured professional knowledge (6); Bookmarking (1); Can answer questions (1)	Provide design exercises (3); Content classification (5); Align comments to the UI image (3)

that it lacked knowledge from external sources ($N = 3$), or felt that the comment summary did not provide sufficient explanation for the quiz ($N = 6$). One participant (PD11, F, 20) also pointed out that she was unable to select her interested UI design examples in the Quizzer.

We get similar comments on the baseline interface to those in experiment I. Participants favored the sense of a community ($N = 6$), the entertaining content in the community ($N = 1$), the broad knowledge view in the comments ($N = 6$), and the diverse UI design examples ($N = 2$). However, they often encountered unrelated comments ($N = 6$) to the knowledge about color and typography, low-quality comments ($N = 2$), or conflicting opinions on the design examples ($N = 1$). Five participants also found that the baseline Reddit interface did not align the comments well to the UI design image, and one participant was unhappy with its messy comment structure. “*I had to scroll back and forth to check the comment and UI design, which is troublesome*” (PB3, F, 22).

Possible long-term usage scenarios. Participants in both groups anticipated that they would use DesignQuizzer ($N = 7$) or the baseline interface ($N = 3$) when they had UI design projects. They would seek design inspiration from the Quizzer ($N = 2$) or the community baseline ($N = 3$). They prefer to study the UI design knowledge with Quizzer ($N = 3$) or baseline ($N = 3$) in their leisure time. One participant (PD1, M, 23) also mentioned that he would use DesignQuizzer when he needed to design a poster, present a PowerPoint, or have painting or Photoshop tasks. Four participants commented that they would like to seek feedback from the design community.

Suggestions for improvement. We asked participants to write down suggestions for improving DesignQuizzer or the community baseline interface especially in their long-term usage. Participants suggested that DesignQuizzer should give feedback to users' designs ($N = 2$), involve more structured knowledge from professional sources ($N = 7$), provide a bookmarking feature ($N = 1$), and be able to answer users' questions on design knowledge ($N = 1$). Users of the baseline interface recommended that it should provide in-situ design exercises that can get feedback from the community ($N = 3$), classify the UI examples and comments ($N = 5$), and adjust the interface to better align comments to the UI images ($N = 3$). These suggestions point out the directions to improve learning support interfaces that leverage the design examples and comments in online communities.

Table 13. (RQ5a) Learned knowledge points about color and typography with DesignQuizzer and the community-like baseline interface in experiment II. The number next to each point is the number of participants who mention it; between-subjects; $N = 28$.

	DesignQuizzer	Baseline
Color	Number (5); Contrast and attention (12); Saturation (1); Palette (7); Specific components (1); Uncommon colors (3); Assignment in the space (2)	Number (1); Contrast and attention (11); Saturation (1); Palette (1); Specific components (4); Uncommon colors (2); Assignment in the space (4)
Typography	Hierarchy (1); Style (5); Size and attention (6)	Style (4); Size and attention (4)

6.3.3 *Learned Knowledge and Application of it in the Visual Design Activity (RQ5).* (RQ5a) Table 13 summarizes the learned knowledge points about color and typography that participants list in the questionnaire. Both user groups reported that they learned visual design knowledge via the explored design examples and comments. For example, they learned that “*an UI should not use a large number of colors; otherwise, it would be a chaos*” (PD9, F, 21). Most participants learned that there should be enough color contrast among the UI components to direct viewers’ attention. “*The UI with a dark background should be careful about the color choices to ensure high readability. The usage of contrasting colors is attractive and informative*” (PD11, F, 20). “*The color contrast should be obvious and allow users to capture the UI’s key information at a glance*” (PB6, M, 20). Besides, the palette can be “*monochromatic*” (PD7, F, 22) with “*gradually varied colors*” (PD10, M, 21). There are also common practices of color usage in specific components, and we’d better not use uncommon colors. “*We should pay attention to the manner reflected by the text color, e.g., do not use red for information such as names of people. Sometimes red can make the UI components stand out, and using contrasting colors can also achieve this effect*” (PD8, F, 20). Last but not least, the color assignment in the UI space should follow principles like “*components of similar types should have the same color*” (PB12, M, 22) and “*there should not be too many colors in the left part and one color in the right part of the UI*” (PD4, F, 21).

The learned knowledge points about typography are mainly about the style and size of fonts. “*The choice of font depends on the UI style. For example, a vivid UI page should not use the boldface that looks serious*” (PD7, F, 22). The size of fonts should direct viewers’ attention to the target area and reflect the hierarchy of the UI page. For instance, “*the text that conveys more important should have larger font size, but it should not be too large; otherwise, it will be stressful*” (PB3, F, 22).

(RQ5b) Figure 7 shows the results regarding changes in participants’ performance on the visual design activity before and after the learning task. For the visual design of a travel agency web page, our results indicate that participants significantly improve their performances in matching the consistent ($F = 21.54, p < 0.001$), distinct ($F = 16.97, p < 0.001$), and intentional ($F = 24.18, p < 0.001$) design principles of color and typography after the learning session. The used learning interface does not significantly impact their performance on this web page design task. However, we observe a significant interaction effect ($F = 5.18, p = 0.031$) between the used interface and time factors on participants’ performance on matching the intentional principle. This indicates that the matched degree with the intentional principle of DesignQuizzer users improved significantly more than the baseline users after the learning session.

For the visual design of a mobile payment page, our results indicate a significant improvement regarding the measure of intention ($F = 5.14, p = 0.032$) in post-test than that in pre-test. There is also a tendency that the measure of consistency improves ($F = 3.89, p = 0.059$) after the learning

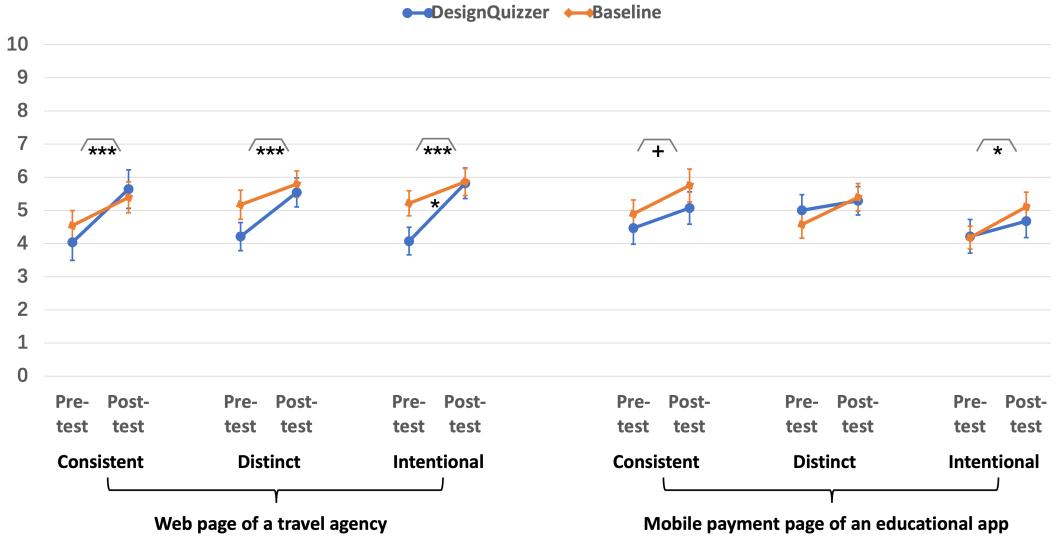


Fig. 7. RQ5b results regarding the changes in participants' performance on the visual design activity before and after the learning session with either DesignQuizzer or baseline interface. In both pre-test and post-test, participants need to color the UI components and adjust the typography of two given uncolored mockups. Two visual design experts rate the designs from aspects of consistency, distinction, and intention, each from 0 to 10 points. *** : $p < 0.001$, * : $p < 0.05$, + : $p < 0.1$.

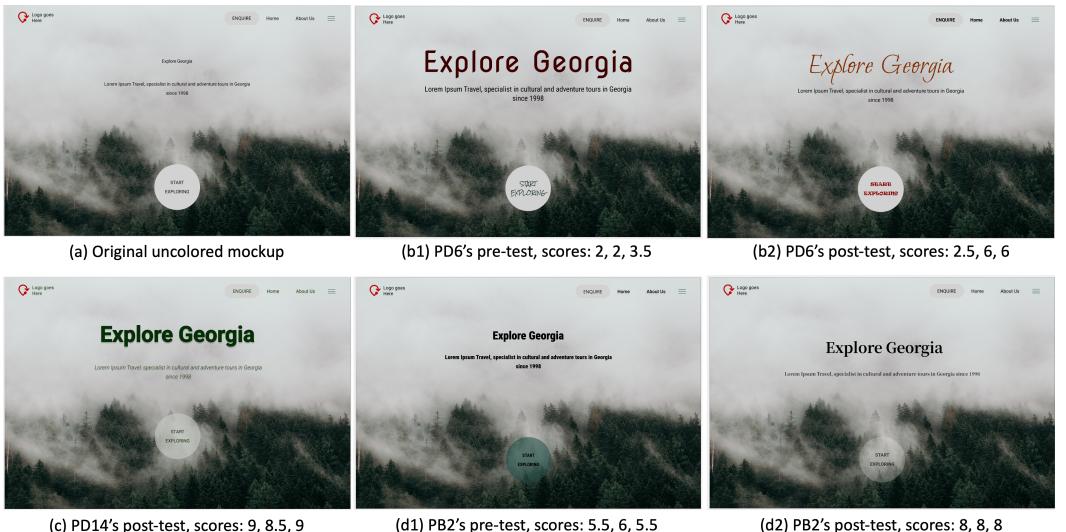


Fig. 8. The original uncolored web page of a travel agency and the sampled outcomes from Participants in DesignQuizzer (note as PD) and Baseline (PB) groups. The average scores given by the two visual design experts are from aspects of consistency, distinction, and intention. This figure is better viewed in color.

session. Neither the used interface nor its interaction with the time factor significantly affects the participants' performance in the visual design activity.

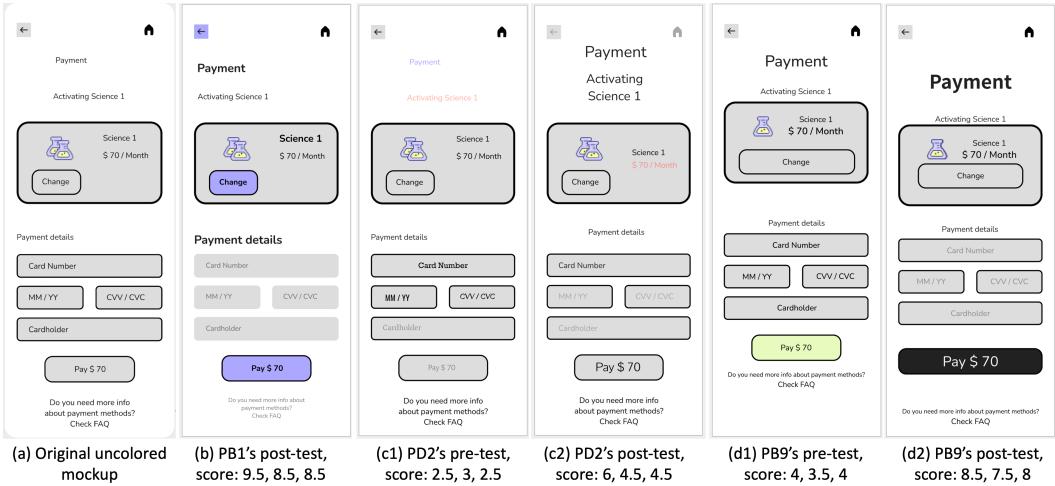


Fig. 9. The original uncolored mobile payment page of an educational app and the sampled outcomes from Participants in DesignQuizzer (note as PD) and Baseline (PB) groups. The average scores given by the two visual design experts are from aspects of consistency, distinction, and intention. This figure is better viewed in color.

Figure 8 and Figure 9 show the sampled outcomes of the visual design activity. For example, PD6 (M, 23) did not perform well in the pre-test of the web page design (Figure 8 b1), as commented by our design expert E1: “*The designer tried to create distinction and hierarchy with purpose. Yet, each component has a unique color and font type, making the page very inconsistent to read. The dark red, the black, and the dark green of the button text are in conflict with each other*”. In the post-test (Figure 8 b2), PD6 improved his performance in distinction and intention, with the red in the “START EXPLORATION” and the bold style in the menu items. PB2 (M, 21) in the baseline group also improved his performance in the post-test of mobile page design (Figure 8 d1 and d2). Our expert E2 left a comment on PB2’s post-test design: “*This UI follows the consistency principle well. For example, the fonts of the text in the same module are consistent. The color system of the clickable color blocks is also united. Besides, the distinction between modules (e.g., the hero title and the description text below) is clear*”.

For the visual design of a mobile payment page, PD2 (M, 21) improved his performance after the learning session (Figure 9 c1 and c2). E2 judged his pre-test design: “*It uses more than four font colors. The font styles are not uniform. What’s worse, there is no distinction between titles and subtitles, the meaning of the spacing is unclear, and the meaning of the operation area is confusing*”. PB9 (M, 20) in the baseline group also applied what he learned in the community to improve his design performance (Figure 9 d1 and d2). E2 commented on his post-test design: “*This UI design clearly distinguishes each module. Its font and color styles are uniform*”.

6.3.4 Visual Design Experts’ Feedback on DesignQuizzer. We interviewed our two visual design experts to obtain their feedback on DesignQuizzer apart from inviting them to score the design outcome. We first walked them through DesignQuizzer’s features. Then, we asked for their comments on the Quizzer’s pros/cons for learning design knowledge and the ways to improve it for novices’ long-term usages. Both experts viewed the DesignQuizzer as a personal design coach for novices and favored its low interaction load. “*It looks like a coach of visual design. The conversation-style design with clickable buttons reduces our interaction effort*” (E2). However, E1 felt that its quizzes

could be too easy for experienced designers. E1 suggested that it should involve professionally-generated content, e.g., the open-sourced design framework and the companies' design library, apart from the current user-generated content in the online design community. Both experts agreed that our classified visual elements and UI components could help users master design knowledge of interests. Nevertheless, to support long-term learning visual design, both experts recommended that DesignQuizzer should have more adaptive learning features. *"It should allow users to archive the UI design examples and knowledge points, such that users can review them in the future. I would also expect that it could recommend related design examples and comments when I click a keyword in the current comment"* (E1).

6.3.5 Summary of the Findings in Experiment II. Our experiment II provides empirical results on its primary RQ5. Both user groups of DesignQuizzer and the community-like baseline learned knowledge points about color and typography via exploring UI design examples and comments. In general, participants in both groups perform better in the visual design activity after the learning session. These results validate the motivation of this paper that the user-generated content in the online design community is promising to support the learning of visual design. We also support the experiment I's findings that the DesignQuizzer is more engaging and perceived as more useful than the baseline interface for learning design knowledge. We additionally collect positive feedback from visual design experts on DesignQuizzer and gain more insights to improve it for novices' long-term usage. In the next section, we discuss the findings from our two experiments, insights for improving community-driven learning support tools, limitations, and future work.

7 DISCUSSION

In this paper, we have presented a community-powered method for learning elementary design concepts interactively with a chatbot. The core of our method is a workflow for extracting and organizing comments into interactive quizzes. Our two evaluation studies support this approach. We found that novices using DesignQuizzer can explore helpful UI examples and comments more efficiently than those using the community-like baseline. Participants attributed these benefits to our quiz pool, which leverages the computational workflow to filter out many lower-quality comments in the community discussion [3, 67, 101, 102].

Our results call for more research to study conversational agents for learning in creative domains. Our approach, which increased the level of engagement and effectiveness in learning, complements the findings of previous works on educational chatbots, like QuizBot for learning factual knowledge [82], Sara for learning programming knowledge [98], and ArgueTutor for learning augmentation writings [93]. After the learning sessions, participants with DesignQuizzer can recognize significantly more points of design principles about the learned visual elements that have or have not been used in unseen design examples.

Our computational approach and the design in DesignQuizzer can serve as starting points for similar approaches in other domains, such as for learning about posters and logos [1, 2] or programming scripts [12]. Researchers can follow our workflow detailed in section 3 in order to generate learning materials. It first requires a labeled dataset regarding the useful content in online communities, e.g., sentences and keywords about a knowledge concept. Researchers can then explore appropriate pre-trained language models and fine-tune them on targeted downstream tasks, e.g., summarization, document/sentence/token classification and clustering, etc. However, we should be aware of the limitations of these computational models. First, they sometimes miss some useful content for learning (e.g., No.6 in Table 2) and make classification mistakes (e.g., No.5 in Table 4). Second, these models usually lack explanations for their results, e.g., what linguistic features would positively contribute to a meaningful rationale sentence. We also offer a promising

way to generate a quiz pool to run a conversational agent based on the materials in an online community. As detailed in subsection 4.2, researchers can first curate the publicly available data online and identify the useful content using the developed computational models above. Next, they should determine the conversation agent’s interaction design and further structure the materials. We adopt a quiz-like design similar to QuizBot [82] and generate the single-choice questions by masking the right answers in the critique sentences. Other possible interaction strategies include 1) arranging a task and providing performance feedback after task completion [93] and 2) prompting open-ended questions a critical thinking tasks [77]. The later open-ended quizzes, as a survey by Wang et al. suggests, would be more valuable compared to the single-choice question [95].

It is also promising to apply our computational approach to powering design support tools, as anticipated by our participants in experiment II. For example, when users want to create a new and unique UI design, they can seek inspiration from the design examples in online communities. Our workflow can help them distill the design examples with constructive feedback and insights on specific visual elements or UI components. Moreover, the workflow can enable a design support tool to help users reflect on their designs. For instance, it can detect the UI components of user designs with computer-vision techniques and retrieve related critiques and suggestions on these components when they appear in similar designs for reference. In a word, to promote reflection and active thinking on users’ designs, a design support tool could adopt a question-driven, conversational design like DesignQuizzer.

Our findings offer two design considerations for improving the effectiveness and user experience of future community-powered learning support tools. First, they should incorporate professional knowledge from external resources into the user-generated learning materials in the communities. In the experiment II, both the DesignQuizzer and baseline user groups were able to apply what they learned to improve their performance in the visual design activity. This finding highlights the value of user-generated content in design communities as the learning materials for novice designers. Nevertheless, the participants in our two experiments pointed out that many comments under UI feedback-request posts often do not provide sufficient explanations for the critiques. As a result, learners can observe what is good or bad in the community-generated learning materials but often do not know why. Therefore, these tools can consider incorporating knowledge graphs/nets (e.g., ConceptNet [60]) and structured documents mined from textbooks or course materials (e.g., [19]) to their knowledge base. Second, the community-powered learning support tools should provide more community-level features for users. For example, there are comments expressing diverse opinions on the design examples or telling jokes, which help to form a social, relaxing, and collaborative learning environment [39]. The learning support tools should present users with this beneficial information, e.g., how many members have similar critiques on this UI example, how many upvotes/likes does the meaningful feedback receive, and how do others agree or disagree with the checked comment. The tools can further offer 1) a “broad” learning mode that encourages users to think from different aspects by showing competing or unrelated comments and 2) a “deep” mode that raises a set of questions on a specific topic (e.g., space) at each learning session.

Our DesignQuizzer can also incorporate the recent chat-tuned large language models (LLMs, e.g., ChatGPT) to power the dialog-based interaction flow. Our work has shown that the rule-based quiz-like dialog flow powered by our computational approaches can achieve good performance in helping users learn visual design. To step forward, we should find ways to fine-tune and control chat-tuned LLMs for DesignQuizzer that could converse with and guide the user more open-endedly. For example, we could design prompts filled by the UI component and visual element keywords to fine-tune LLMs for generating related questions and providing feedback. Moreover, the advanced large multimodal model (e.g., GPT4 [73]) can help to align the comments to the UI image from online communities (Table 12). For instance, we could select a detected UI component in the comment

and query GPT4 to highlight the related area of the component in the UI image, which can help users easily locate their interested area.

Our work has several limitations that call for future work. First, while our computational models have acceptable performance in terms of automatic evaluation metrics, they sometimes make mistakes (e.g., [Table 4](#)), and they can not explicitly identify the relevancy among the critique, rationale, and suggestion sentences. Although our participants did not report these issues in the experiment, future work should improve our computational models with more training data. We can iteratively enlarge the training dataset by manually refining the machine-generated labels and re-train the models with the updated dataset. Second, we did not have a study separating the impact of organized comments and the quiz-like interaction design on our measured items. The qualitative feedback in the interviews in experiment I suggests that the organized information contributed most to improving efficiency on exploring helpful examples and comments, and that the active question answering helped most with engagement. However, understanding the precise contributions of each system feature remains an important issue for future work. Third, we evaluated the tools' impact on learners' user experience and outcome in two lab studies but not a longer-term study. Thus, our empirical results can not expel the novelty effects of people using our DesignQuizzer for the first time. Fourth, our participants are novice designers who have little or no experience in learning UI visual design. It would be interesting to examine how regular designers can learn with DesignQuizzer and the online design community. For future work, we suggest measuring the long-term learning experience and effectiveness with users of diverse backgrounds.

8 CONCLUSION

In this paper, we facilitate novices to learn visual elements in online communities by proposing a computational approach for organizing comments and a conversational agent DesignQuizzer. We presented methods to extract meaningful feedback from comments on UI designs, classify the feedback into critique, suggestion, and rationale sentences, recognize the keywords about visual elements and UI components in the sentences, and cluster the keywords into groups of higher-level concepts. This approach allows DesignQuizzer's quiz-like interaction with users. We compared DesignQuizzer with a community-like baseline interface in a within-subjects experiment I with 24 novices and a between-subjects experiment II with 28 novices. The results show that DesignQuizzer significantly improves participants' efficiency and engagement in learning visual design from examples and comments in the communities. Participants rated the Quizzer significantly more useful and easier to use for their learning tasks, and they favored its questions for promoting active thinking. Participants with DesignQuizzer can apply what they have learned to criticize others' UI designs and enhance their performance in the visual design activity. We also offered insights for generalizing our approach to other online communities and provided design considerations for improving the effectiveness and user experience of community-powered learning support tools. We hope our work will attract more researchers to leverage the resources in online communities to build intelligent learning support systems.

ACKNOWLEDGMENTS

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China with Grant No. 62202509 and partially supported by the Research Grants Council of the Hong Kong Special Administrative Region under General Research Fund (GRF) with Grant No. 16203421. Antti Oulasvirta was supported by the Research Council of Finland (flagship program: Finnish Center for Artificial Intelligence, FCAI, grants 328400, 345604, 341763; Human Automata, grant 328813).

REFERENCES

- [1] 2008. Reddit r/graphic_design. Accessed in September, 2022 from https://www.reddit.com/r/graphic_design/.
- [2] 2010. Reddit r/design_critiques. Accessed in September, 2022 from https://www.reddit.com/r/design_critiques/.
- [3] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. 183–194.
- [4] Louis Alfieri, Timothy J. Nokes-Malach, and Christian D. Schunn. 2013. Learning Through Case Comparisons: A Meta-Analytic Review. *Educational Psychologist* 48 (2013), 113 – 87.
- [5] Robert K. Atkinson, Sharon J. Derry, Alexander Renkl, and Donald W. Wortham. 2000. Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research* 70 (2000), 181 – 214.
- [6] Sasha A Barab and Thomas Duffy. 2012. From practice fields to communities of practice. In *Theoretical foundations of learning environments*. Routledge, 29–65.
- [7] Terry Barrett. 1988. A Comparison of The Goals of Studio Professors Conducting Critiques and Art Education Goals for Teaching Criticism.
- [8] Julie Campbell, Cecilia Aragon, Katie Davis, Sarah Evans, Abigail Evans, and David Randall. 2016. Thousands of Positive Reviews: Distributed Mentoring in Online Fan Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) (CSCW '16). Association for Computing Machinery, New York, NY, USA, 691–704. <https://doi.org/10.1145/2818048.2819934>
- [9] Julie Campbell, Cecilia Aragon, Katie Davis, Sarah Evans, Abigail Evans, and David Randall. 2016. Thousands of positive reviews: Distributed mentoring in online fan communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 691–704.
- [10] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.
- [11] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. <https://doi.org/10.1080/10447318.2020.1841438> arXiv:<https://doi.org/10.1080/10447318.2020.1841438>
- [12] Ruijia Cheng, Sayamindu Dasgupta, and Benjamin Mako Hill. 2022. How Interest-Driven Content Creation Shapes Opportunities for Informal Learning in Scratch: A Case Study on Novices' Use of Data Structures. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 228, 16 pages. <https://doi.org/10.1145/3491102.3502124>
- [13] Ruijia Cheng and Benjamin Mako Hill. 2022. Many Destinations, Many Pathways: A Quantitative Analysis of Legitimate Peripheral Participation in Scratch. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 381 (nov 2022), 26 pages. <https://doi.org/10.1145/3555106>
- [14] Ruijia Cheng and Benjamin Mako Hill. 2022. Many Destinations, Many Pathways: A Quantitative Analysis of Legitimate Peripheral Participation in Scratch. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 381 (nov 2022), 26 pages. <https://doi.org/10.1145/3555106>
- [15] Ruijia Cheng and Mark Zachry. 2020. Building Community Knowledge In Online Competitions: Motivation, Practices and Challenges. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 179 (oct 2020), 22 pages. <https://doi.org/10.1145/3415250>
- [16] Ruijia Cheng, Ziwen Zeng, Maysnow Liu, and Steven Dow. 2020. Critique Me: Exploring How Creators Publicly Request Feedback in an Online Critique Community. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 161 (oct 2020), 24 pages. <https://doi.org/10.1145/3415232>
- [17] Ruijia Cheng, Ziwen Zeng, Maysnow Liu, and Steven Dow. 2020. Critique Me: Exploring How Creators Publicly Request Feedback in an Online Critique Community. *Proceedings of the ACM on Human-Computer Interaction* 4, Article 161 (2020), 24 pages. Issue CSCW2.
- [18] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. <https://doi.org/10.1080/00461520.2014.965823> arXiv:<https://doi.org/10.1080/00461520.2014.965823>
- [19] Jerry Cao Chris Bank. [n.d.]. Web UI Design Best Practices: UI Design From The Experts. Accessed on September 2022, <https://www.uxpinc.com/studio/ebooks/web-ui-design-best-practices/>.
- [20] Patrick A Crain and Brian P Bailey. 2017. Share Once or Share Often? Exploring How Designers Approach Iteration in a Large Online Community. ACM, New York, NY, USA, 80–92.
- [21] Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row, New York, USA.
- [22] Sayamindu Dasgupta, William Hale, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Remixing as a pathway to computational thinking. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1438–1449.

- [23] Sayamindu Dasgupta and Benjamin Mako Hill. 2018. How “wide walls” can increase engagement: evidence from a natural experiment in Scratch. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [25] Rahul R Divekar, Haley Lepp, Pravin Chopade, Aaron Albin, Daniel Brenner, and Vikram Ramnarayanan. 2021. Conversational Agents in Language Education: Where They Fit and Their Research Challenges. In *International Conference on Human-Computer Interaction*. Springer, 272–279.
- [26] doccano. 2022. Text Annotation for Humans. Accessed in September, 2022 from <https://doccano.herokuapp.com/>.
- [27] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1013–1022.
- [28] Hugging Face. 2022. Sentence Transformers. Accessed in September, 2022 from <https://huggingface.co/sentence-transformers>.
- [29] Hugging Face. 2022. Summarization. Accessed in September, 2022 from <https://huggingface.co/docs/transformers/tasks/summarization>.
- [30] Hugging Face. 2022. Token Classification. Accessed in September, 2022 from <https://huggingface.co/course/chapter7/2?fw=pt>.
- [31] Edmund Burke Feldman. 1994. *Practical art criticism*. Pearson.
- [32] Eureka Foong, Steven P. Dow, Brian P. Bailey, and Elizabeth M. Gerber. 2017. Online Feedback Exchange: A Framework for Understanding the Socio-Psychological Factors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (*CHI ’17*). Association for Computing Machinery, New York, NY, USA, 4454–4467. <https://doi.org/10.1145/3025453.3025791>
- [33] Denae Ford, Kristina Lustig, Jeremy Banks, and Chris Parnin. 2018. “We Don’t Do That Here”: How Collaborative Editing with Mentors Improves Engagement in Social Q&A Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI ’18*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3174182>
- [34] John Frens, Erin Walker, and Gary Hsieh. 2018. Supporting answerers with feedback in social Q&A. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 1–10.
- [35] Emilia F. Gan, Benjamin Mako Hill, and Sayamindu Dasgupta. 2018. Gender, Feedback, and Learners’ Decisions to Share Their Creative Computing Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 54 (nov 2018), 23 pages. <https://doi.org/10.1145/3274323>
- [36] Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and Transfer: A General Role for Analogical Encoding. *Journal of Educational Psychology* 95 (2003), 393–408.
- [37] Google. 2022. Material Design. Accessed in September, 2022 from <https://material.io/components>.
- [38] Ben Gorvine, Karl Rosengren, Lisa Stein, and Kevin Biolsi. 2017. *Research Methods: From Theory to Practice - Chapter 14: Analyzing Your Data II: Specific Approaches Inside Research*. Oxford University Press. <https://global.oup.com/us/companion.websites/9780190201821/sr/outline/ch14/>
- [39] Colin M Gray and Yubo Kou. 2019. Co-producing, curating, and defining design knowledge in an online practitioner community. *CoDesign* 15, 1 (2019), 41–58.
- [40] Aaron Halfaker, Os Keyes, and Dario Taraborelli. 2013. Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) (*CSCW ’13*). Association for Computing Machinery, New York, NY, USA, 849–860. <https://doi.org/10.1145/2441776.2441872>
- [41] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [42] Lena Hegemann, Niraj Ramesh Dayama, Abhishek Iyer, Erfan Farhadi, Ekaterina Marchenko, and Antti Oulasvirta. 2023. CoColor: Interactive Exploration of Color Designs. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (*IUI ’23*). Association for Computing Machinery, New York, NY, USA, 106–127. <https://doi.org/10.1145/3581641.3584089>
- [43] Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. 2009. Getting Inspired! Understanding How and Why Examples Are Used in Creative Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (*CHI ’09*). Association for Computing Machinery, New York, NY, USA, 87–96. <https://doi.org/10.1145/1518701.1518717>

- [44] Julie S Hui, Elizabeth M Gerber, and Steven P Dow. 2014. Crowd-based design activities: helping students connect with users online. In *Proceedings of the 2014 conference on Designing Interactive Systems*. 875–884.
- [45] Hyeyoun B Kang, Gabriel Amoako, Neil Sengupta, and Steven P Dow. 2018. Paragon: An online gallery for enhancing design feedback with visual examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [46] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting Visual Metaphor Ideation through Multi-Dimensional Example-Based Exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 427, 15 pages. <https://doi.org/10.1145/3411764.3445325>
- [47] Mostafa Keikha, Jae Hyun Park, and W Bruce Croft. 2014. Evaluating answer passages using summarization measures. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 963–966.
- [48] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the Web with Natural Language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI ’22). Association for Computing Machinery, New York, NY, USA, Article 5, 17 pages. <https://doi.org/10.1145/3491102.3501931>
- [49] Janin Koch, Magda Laszlo, Andres Lucero, and Antti Oulasvirta. 2018. Surfing for Inspiration: digital inspirational material in design practice. In *Design Research Society International Conference*. Design Research Society, 1247–1260.
- [50] Yubo Kou and Colin M Gray. 2017. Supporting distributed critique through interpretation and sense-making in an online creative community. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–18.
- [51] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M. Gerber, Brian P. Bailey, and Steven P. Dow. 2017. Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI ’17). Association for Computing Machinery, New York, NY, USA, 4627–4639. <https://doi.org/10.1145/3025453.3025883>
- [52] Sneha R. Krishna Kumaran, Deana C. McDonagh, and Brian P. Bailey. 2017. Increasing Quality and Involvement in Online Peer Feedback Exchange. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 63 (dec 2017), 18 pages. <https://doi.org/10.1145/3134698>
- [53] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 20, 6, Article 33 (dec 2013), 31 pages. <https://doi.org/10.1145/2505057>
- [54] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 75–84.
- [55] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
- [56] Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R. Klemmer. 2010. Designing with Interactive Example Galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI ’10). Association for Computing Machinery, New York, NY, USA, 2257–2266. <https://doi.org/10.1145/1753326.1753667>
- [57] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3313831.3376590>
- [58] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [59] Chengzhong Liu, Zeyu Huang, Dingdong Liu, Shixu Zhou, Zhenhui Peng, and Xiaojuan Ma. 2022. PlanHelper: Supporting Activity Plan Construction with Answer Posts in Community-Based QA Platforms. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 454 (nov 2022), 26 pages. <https://doi.org/10.1145/3555555>
- [60] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.
- [61] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 <http://arxiv.org/abs/1907.11692>
- [62] Nichola Lubold, Erin Walker, Heather Pon-Barry, and Amy Ogan. 2018. Automated pitch convergence improves learning in a social, teachable robot for middle school mathematics. In *International conference on artificial intelligence in education*. Springer, 282–296.
- [63] Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer’s own writing. *Journal of Second Language Writing* 18 (2009), 30–43.
- [64] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2015. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th*

ACM conference on computer supported cooperative work & social computing. 473–485.

- [65] Anderson LW, Krathwohl DR, Airasian PW, Cruikshank KA, Richard Mayer, Pintrich PR, J. Raths, and Wittrock MC. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Pearson; 1st edition.
- [66] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
- [67] Jennifer Marlow and Laura Dabbish. 2014. From rookie to all-star: professional development in a graphic design social networking site. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 922–933.
- [68] MasterClass. 2021. Elements of Design: Understanding the 7 Elements of Design. Accessed in September, 2022 from <https://www.masterclass.com/articles/elements-of-design-explained>.
- [69] Mark W. Newman and James A. Landay. 2000. Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York City, New York, USA) (*DIS '00*). Association for Computing Machinery, New York, NY, USA, 263–274. <https://doi.org/10.1145/347642.347758>
- [70] Tricia J. Ngoon, C. Ailie Fraser, Ariel S. Weingarten, Mira Dontcheva, and Scott Klemmer. 2018. *Interactive Guidance Techniques for Improving Creative Feedback*. Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3173574.3173629>
- [71] Benjamin Nye, Arthur Graesser, and Xiangen Hu. 2014. AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education* 24 (12 2014). <https://doi.org/10.1007/s40593-014-0029-5>
- [72] Heather O'Brien. 2016. *Theoretical Perspectives on User Engagement*. Springer International Publishing, Cham, 1–26. https://doi.org/10.1007/978-3-319-27446-1_1
- [73] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [74] Jonas Oppenlaender, Elina Kuosmanen, Andrés Lucero, and Simo Hosio. 2021. Hardhats and Bungaloos: Comparing Crowdsourced Design Feedback with Peer Design Feedback in the Classroom. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 570, 14 pages. <https://doi.org/10.1145/3411764.3445380>
- [75] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '20*). ACM, New York, NY, USA, 556–567. <https://doi.org/10.1145/3313831.3376695>
- [76] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300328>
- [77] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. 2022. CReBot: Exploring interactive question prompts for critical paper reading. *International Journal of Human-Computer Studies* 167 (2022), 102898.
- [78] Zhenhui Peng, Xiaojuan Ma, Diyi Yang, Ka Wing Tsang, and Qingyu Guo. 2021. Effects of Support-Seekers' Community Knowledge on Their Expressed Satisfaction with the Received Comments in Mental Health Communities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 536, 12 pages. <https://doi.org/10.1145/3411764.3445446>
- [79] Pushshift. 2022. Reddit Statistics - pushshift.io. Accessed in September, 2022 from <https://pushshift.io/>.
- [80] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharhan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. <http://jmlr.org/papers/v21/20-074.html>
- [81] Mitchel Resnick, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay Silver, Brian Silverman, and Yasmin Kafai. 2009. Scratch: Programming for All. *Commun. ACM* 52, 11 (nov 2009), 60–67. <https://doi.org/10.1145/1592761.1592779>
- [82] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300587>
- [83] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
- [84] Christopher Scaffidi and Christopher Chambers. 2012. Skill progression demonstrated by users in the Scratch animation environment. *International Journal of Human-Computer Interaction* 28, 6 (2012), 383–398.

- [85] Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [86] Samantha Shorey, Benjamin Mako Hill, and Samuel Woolley. 2021. From hanging out to figuring it out: Socializing online as a pathway to computational thinking. *New Media & Society* 23, 8 (2021), 2327–2344.
- [87] Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing answers in non-factoid community question-answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 405–414.
- [88] Rohail Syed, Kevyn Collins-Thompson, Paul Bennett, Mengqiu Tang, Shane Williams, Shamsi Iqbal, and Wendy Tay. 2020. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *The Web Conference 2020 (formerly WWW conference)*. <https://www.microsoft.com/en-us/research/publication/improving-learning-outcomes-with-gaze-tracking-and-automatic-question-generation/>
- [89] Yla Tausczik and Ping Wang. 2017. To Share, or Not to Share? Community-Level Collaboration in Open Innovation Contests. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 100 (dec 2017), 23 pages. <https://doi.org/10.1145/3134735>
- [90] Yla R Tausczik, Aniket Kittur, and Robert E Kraut. 2014. Collaborative problem solving: A study of mathoverflow. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 355–367.
- [91] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and analysis of self-disclosure in online news commentaries. In *The World Wide Web Conference*. 3272–3278.
- [92] Viswanath Venkatesh and Hillol Bala. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences* 39, 2 (2008), 273–315. <https://doi.org/10.1111/j.1540-5915.2008.00192.x> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5915.2008.00192.x>
- [93] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. *ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411764.3445781>
- [94] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A conversational agent to improve response quality in course evaluations. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–9.
- [95] Xu Wang, Carolyn Rose, and Ken Koedinger. 2021. Seeing beyond expert blind spots: Online learning design for scale and quality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [96] Florian Weber, Thiemo Wambsganß, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education Completed Research Paper.
- [97] Nathaniel Weinman, Steven M. Drucker, Titus Barik, and Robert DeLine. 2021. Fork It: Supporting Stateful Alternatives in Computational Notebooks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 307, 12 pages. <https://doi.org/10.1145/3411764.3445527>
- [98] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI ’20). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376781>
- [99] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.
- [100] Robert Woolson. 2008. Wilcoxon Signed-Rank Test. <https://doi.org/10.1002/9780471462422.eoct979>
- [101] Anbang Xu and Brian Bailey. 2012. What do you think? A case study of benefit, expectation, and interaction in a large online critique community. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. 295–304.
- [102] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (Baltimore, Maryland, USA) (CSCW ’14). Association for Computing Machinery, New York, NY, USA, 1433–1444. <https://doi.org/10.1145/2531602.2531604>
- [103] Litao Yan, Elena L. Glassman, and Tianyi Zhang. 2021. Visualizing Examples of Deep Neural Networks at Scale. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 313, 14 pages. <https://doi.org/10.1145/3411764.3445654>
- [104] Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI ’19). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300574>

- [105] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–15.
- [106] Seungwon Yang, Carlotta Domeniconi, Matt Revelle, Mack Sweeney, Ben U Gelman, Chris Beckley, and Aditya Johri. 2015. Uncovering trajectories of informal learning in large online communities of creators. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. 131–140.
- [107] Yu-Chun Grace Yen and Steven P. Dow. 2022. Seeking Exemplars in the Wild: Exploring How Students Find Design Examples to Support Personalized Learning. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) (*L@S '22*). Association for Computing Machinery, New York, NY, USA, 418–421. <https://doi.org/10.1145/3491140.3528303>
- [108] Yu-Chun Grace Yen, Joy O Kim, and Brian P Bailey. 2020. Decipher: an interactive visualization tool for interpreting unstructured design feedback from multiple providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [109] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1005–1017.

Received January 2023; revised October 2023; accepted December 2023