"I'd be watching him contour till 10 o'clock at night": Understanding Tensions between Teaching Methods and Learning Needs in Healthcare Apprenticeship

Matin Yarmand
UC San Diego, The Design Lab
United States
myarmand@ucsd.edu

Chen Chen
UC San Diego, The Design Lab
United States
chenchen@ucsd.edu

Kexin Cheng
UC San Diego, Cognitive Science
United States
k5cheng@ucsd.edu

James D. Murphy
UC San Diego, Radiation Medicine
United States
j2murphy@ucsd.edu

Nadir Weibel
UC San Diego, The Design Lab
United States
weibel@ucsd.edu

ABSTRACT

Apprenticeship is the predominant method for transferring specialized medical skills, yet the inter-dynamics between faculty and residents, including methods of feedback exchange are under-explored. We specifically investigate contouring: outlining tumors in preparation for radiotherapy, a critical skill that when performed subpar, severely degrades patient survival. Interviews and design-thinking workshops (N = four faculty; six residents) revealed misalignment between teaching methods and residents who desired timely, relevant, and diverse feedback. We further discuss reasons: overlapping learning content and strategies to ease tensions between clinical and teaching duties, and lack of support for exchange of cognitive processes. The follow-up survey study (N = 67 practitioners from 31 countries), which contained annotation and sketching tasks, provided diverse perspective over effective feedback elements. We lastly present sociotechnical implications in supporting faculty's teaching duties and learners' cognitive models, such as systematically leveraging senior learners in providing case-based guidance and supporting double-sided flow of cognitive information via in-situ video snippets.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Contouring, Healthcare Training, Cognitive Apprenticeship

ACM Reference Format:
Matin Yarmand, Chen Chen, Kexin Cheng, James D. Murphy, and Nadir Weibel. 2024. "I'd be watching him contour till 10 o'clock at night": Understanding Tensions between Teaching Methods and Learning Needs in Healthcare Apprenticeship. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642453

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642453

1 INTRODUCTION

Apprenticeship models of training have been key in transmitting specialized knowledge and skills from experts to novices, particularly in critical domains that involve complex cognitive processes and require high quality task completion, such as healthcare. Apprenticeship refers to direct observation and supervision between learners and experts until the apprentice is proficient enough to accomplish the task independently. While *traditional* apprenticeship involves learning a physical and tangible activity, many specialized practices contain less visible, yet cognitively complex tasks. *Cognitive* apprenticeship [14] is a model that aims to make internal cognitive models more visible by following six principles of learning: modeling, coaching, scaffolding, articulation, reflection, and exploration. Healthcare is a domain that contains many high-stakes, specialized, and cognitively-complex tasks, in which cognitive apprenticeship is particularly suited as the predominant training model. Medical residency programs, and specifically the task of contouring in radiation oncology, is a unique case in that despite relying on apprenticeship methods of teaching, is prone to detrimental mistakes which can stem from scarce availability of expert faculty and subpar training methods. Contouring is a high-stakes task that refers to the identification of tumor and organs at risk during the radiation treatment planning process. Poor radiation planning occurs at a large scale and leads to detrimental consequences for patient well-being. Over- and under-contoured plans lead to excess toxicity to the nearby healthy organs, or insufficient radiation to the tumorous cells which will increase the risk of disease recurrence. Clinical trials reveal that protocol violations — which can occur up to a staggering 81% of radiation plans [39] — can decrease patient survival by 22% [94]. Given how radiation oncology faculty possess a dual role of clinician and teacher, when availability is limited, the clinician role takes absolute priority over teaching duties, potentially contributing to a subpar apprenticeship model of training. As such, it is imperative to understand the existing mechanisms of contouring education and examine the dynamics of feedback exchange between the faculty and residents in the apprenticeship model of residency programs.

This paper explores the dynamics between faculty and residents in healthcare apprenticeship, and especially the methods of feedback exchange in the transfer of contouring skills in radiation oncology. Interviews with four faculty and six residents identified existing training strategies and revealed residents' perceptions, such as 1-on-1 contouring watch-alongs — *i.e.,* when a faculty contours an entire case and thinks out-loud their processes as the resident watches — which residents found tedious and marginally beneficial. Instead, learners emphasized the importance of timely, targeted, and diverse feedback, as revealed by two design-thinking workshops in which participants designed their ideal contouring feedback interfaces. The created designs later shaped the content of a reflective-style survey study that aimed to assess the effectiveness of granular elements of these interfaces given a diverse population of physicians, including 67 practitioners from 31 countries. We discuss three socio-technical findings arising from our studies that have implications not just for contouring education, but also for broader healthcare apprenticeship models:

(1) we note that the faculty's dual role of clinician and teacher leads to the design of learning content and strategies that are not fully aligned with the learners' skill-level, but aim to satisfy clinical duties at the same time.

(2) we report on how healthcare apprenticeship aligns more closely with a traditional model, and lacks effective support for articulation, reflection, and exploration of a cognitive apprenticeship model.

(3) we propose practical sociotechnical solutions that aim to mitigate points (1) and (2), such as, leveraging peer resident resources, and aggregating variability and promoting deliberation.

The findings from this paper contribute to a multi-faceted understanding of healthcare residency programs via the cognitive apprenticeship model, and further offers key sociotechnical considerations for introducing computer-supported training tools in healthcare.

2 BACKGROUND AND RELATED WORK

This section describes cognitive apprenticeship and Human-Computer Interaction (HCI) systems that support this model of training, contouring process, and importance of user interface design to support healthcare training.

2.1 Curricular and Technological Support for Facilitating Cognitive Apprenticeship

While *traditional* apprenticeship is an effective instructional model for transferring physical skills from on-site supervision of an expert, *cognitive* apprenticeship [12, 13] focuses on developing stronger mental models and metacognitive skills, especially in tasks that are not fully observable [6, 43]. In other words, cognitive apprenticeship elevates the precursory model by making the tacit knowledge of experts explicit [81] using a six-step principle, as defined in Table 1: modeling, coaching, scaffolding (which comprise the traditional model), followed by articulation, reflection, and exploration [13]. Broadly, the first three steps are the core principles of traditional apprenticeship. The additional *Articulation* and *Reflection* steps aim to highlight the expert's model of problem-solving, and also encourage learners to gain control of their own problem-solving strategies. The last step (*i.e., Exploration*) fosters learner anatomy, not just in terms of problem-solving, but also problem-setting.

Table 1: The six principles of cognitive apprenticeship, formulated and defined by Collins et al. [12]. The first three principles comprise the traditional model of apprenticeship.

Principle	Definition
Modeling	Expert performs specialized task and externalizes internal processes and activities, while learner observes.
Coaching	Learner performs specialized task, while expert observes and offers feedback, including hints and reminders.
Scaffolding	Expert diagnoses learner's skill level and task difficulty, and adjusts time and content of feedback accordingly.
Articulation	Learner articulates their knowledge, reasoning, and internal processes, while expert assesses learner's understanding.
Reflection	Learner compares their problem-solving processes with a cognitive model of expertise involving processes of expert or peer learners.
Exploration	Expert encourages learner to pursue and solve new problems independently by setting relevant learning goals.

Many educational programs offer heuristic strategies and logistical support to implement cognitive apprenticeship in different learning tasks, such as reading [60, 61], writing [71], multimedia design [52], high school science [68], college math [74, 75], doctoral research methods [21], and healthcare [6, 69]. A primary principle for these methods is to guide learners to think through and solve problems similarly to how an expert approaches it: For instance, Scardamalia et al. [71] construct a sophisticated set of procedural heuristics according to novices' "knowledge-telling" *v.s.* experts' "knowledge tansforming" [70]: while novice writers tend to immediately produce text by writing down ideas sequentially, experts spend time not only on writing, but also planning and revising a cohesive story. Healthcare research has also explored and implemented cognitive apprenticeship strategies in different contexts such as psychiatric nursing college [45], trauma life support course in a medical school [18], and junior radiology residency curricula [88]. Given the need for teaching specialized medical skills in high-stakes clinical domains, more research is key to capture the intricacies of different fields and potentially contribute to a holistic understanding of cognitive apprenticeship in healthcare. This work sheds light on dynamics of the existing apprenticeship model training (in the case of contouring in radiation oncology) and reveals a lack of support for developing the internal cognitive models of learners.

In addition to instructional programs, HCI and Educational Computing literature further introduced computer-supported tools to support apprenticeship [83, 84, 89]. To improve the scale of

apprenticeship among crowdworkers, Suzuki et al. [84] introduced *Atelier* which matched less experienced workers (*i.e.,* mentees) with others who are more skilled (*i.e.,* mentors) and facilitated micro-internships as the mentee completed real-world tasks and received feedback from the assigned mentor. Cognitive Apprenticeship Web-based Argumentation (CAWA) [89] aimed to facilitate cognitive apprenticeship in large classroom settings by providing individualized assistance in articulating, reflecting, and exploring skills related to argumentation, an important component in STEM education. Yin *et al.* [83] developed a system that addresses an important limitation of apprenticeship in endodontic surgery: assessing the practice outcome (in a virtual reality simulation) and providing formative and individualized feedback. In healthcare, given the physician experts' dual role of clinician and teacher, patient care takes absolute priority over teaching [66]. As such, computer-supported tools that provide adaptive and timely feedback can enhance the overall cognitive apprenticeship and lead to better medical training and patient outcome. Following design-thinking workshops and reflective-style survey studies, this work explores effective feedback elements of computerized support that can mitigate pedagogical duties of faculty while enhancing learning experience of residents.

2.2 Contouring: Background and Learning Resources in Residency Programs

Radiation oncologists perform contouring – using desktop based softwares such as MIM[1] and Eclipse[2] – by repeatedly drawing 2D contours on relevant image slices to encompass the 3D volume of the tumorous tissues. While the final contours on CT scans influence dose calculation, different types of images and planes can inform decision-making: for instance, physicians use MRI images to treat brain cancer, because brain organs appear more distinctly in these scans compared to CT images. The oncologists can also consult different orientations of the same set of images to inform anatomy of structures.

Contouring is considered the weakest link in radiation oncology treatment [58] due to substantial variability in providers' contours [26] and mistakes that lead to detrimental consequences for patient safety and survival. Radiation plans that deviate from protocol specifications substantially decrease survival compared to patients with compliant radiation plans: for instance, two clinical trials in head-and-neck cancer revealed 20% and 22% decrease in survival due to protocol violations [63, 94]. In addition, clinical trials reveal sobering insights into the high frequency of poor contouring: a study on anal cancer found that 81% of radiation plans had "incorrect contours" [39] and 70% of contours on brain cancer cases were "unacceptable" [22].

While auxiliary educational resources (*e.g.,* atlases) and emergent virtual reality tools [8, 9] can improve contouring skills, direct learning from the attending faculty remains as the main method of training in residency programs. Medical reference aids (*e.g.,* atlases and books [30, 47]) can mitigate the existing variability and improve contour agreement [15]. In practice, however, sub-optimal methods of development, delivery, and access hinder potential benefits from these resources [31–33]. One strategy to improve access to contouring guidelines is web-based 3D atlases: as an example, eContour[3] [77] is a browser-based atlas that can improve contouring accuracy and anatomy knowledge [27], and further demonstrated higher usability and learnability [62]. Recent works explored cross-device and on-demand feedback strategies in terms of percentage of overlap with expert contours and step-by-step guidance on regions of interest [98, 99]. Despite the existing medical reference aids, receiving one-to-one supervision from the faculty (in an apprenticeship model [73]) remains the main method of training in contouring education, as also seen in many other residency programs (*e.g.,* psychiatry, surgery, and radiology) [23]. Residency programs in radiation oncology assign residents to one expert faculty at a time (*a.k.a.* attending physician) with residents learning contouring practices by observing the faculty's general workflow and re-creating their processes. This work aims to improve contouring education by examining dynamics of feedback exchange between radiation oncology faculty and residents, and further offering practical sociotechnical solutions.

2.3 Impact of User Interface Design on Decision-making and Training in Healthcare

Many healthcare-focused HCI research investigated improving tools and interfaces used by single clinicians, while many CSCW papers in medical domains outlined problems and opportunities for designing interfaces that foster collaboration in clinical teams, with some recent works exploring Human-AI interaction in diagnostic settings. This section provides a brief overview of the relevant HCI and CSCW research, and situates this work (and the broader healthcare training) in the existing literature.

Starting with the work of Grudin in the late 80s [35], a considerable number of the HCI and CSCW literature focused on understanding why applications built for collaboration in the workplace fail to achieve their goals. Grudin attributed the lack of contextual research [92] to this failure, and Ehn and Kyng [19] advocated for better understanding of the stakeholders by "working beside them a long time in order to develop a new system that is *owned* by the workers". Building on this research, Markus and Connolly [55] argued that the adoption of tools that are used in a multi-user setting in the workplace heavily depends on the interdependence in the payoffs of different users. To understand these tensions in healthcare — in which multiple stakeholders need to engage in decision-making and agree on terms that will lead to life or death outcomes — more recently Schaekermann *et al.* [72] studied factors that lead to experts' disagreements and their justifications, and how the presentation of the data is key to engage in effective decision-making. This is true in terms of both medical time series data [72], but even more importantly when data have a higher degree of interpretability such as in medical image-based comparison [7, 95]. Specifically, Cai *et al.* [7] outlined how tools in the context of image retrieval systems for medical decision-making need to facilitate interaction across clinicians and AI-aids, in such a way that clinical teams can trust and effectively

[1] MIM Maestro: https://www.mimsoftware.com/radiationoncology/maestro.
[2] Eclipse: https://www.varian.com/products/radiotherapy/treatment-planning/eclipse.
[3] eContour: https://econtour.org.

Table 2: Background details on the four faculty and six residents who participated in this study.

ID	Title	Gender	Age	Experience
F1	Assistant Professor	Male	32	5 years
F2	Assistant Professor	Female	39	11 years
F3	Assistant Professor	Male	34	6 years
F4	Assistant Professor	Male	37	7 years
R1	Resident Year 4	Male	35	4 years
R2	Resident Year 4	Female	33	4 years
R3	Resident Year 1	Female	28	6 months
R4	Resident Year 2	Male	29	1 years
R5	Resident Year 2	Male	29	1 years
R6	Resident Year 3	Male	33	2 years

work with this data. Xie *et al.* [95] explored a similar setting and highlighted the importance of designing tools that can support clinical teams to engage in AI-enabled chest X-ray analysis.

While these HCI and CSCW works (among others) are key to advancing the effective development of interfaces for the practice of medicine, a large part of the clinical experience involves training medical students and residents. Learning how to use these systems and interfaces is an important part of the learning experience, but very often, the same tools that are good at delivering care, have not been designed to support training and effective decision-making for trainees. As laid out by Markus and Connolly [55], to make an interface successful, we need to look at the interdependence in the payoffs of the different users, and one of the users in this case is a trainee (*e.g.,* resident) who learns from the expert (*e.g.,* attending faculty). While there is a lack of research specifically in healthcare training, prior works in other educational settings showed benefits of careful interface design for learners: for example, recent work showed how particular user interface add-ons can alleviate confusion and enable learners to better understand the expert content communicated to them [97], and how referencing back to material that the learner previously engaged with increases satisfaction and results in more effective learning [100].

This paper takes a user-centered design approach that aims to surface similar paradigms in the context of healthcare training, specifically for the case of image-based comparison and radiation oncology. After careful examination of the context and defining the existing interrelationships between residents and faculty, this work first explores effective feedback design elements, and later offers practical sociotechnical guidelines that improve contouring education, and more broadly, healthcare apprenticeship.

3 METHODS

This study followed a two-step user-centered design protocol. Through the official residency mailing list of the Department of Radiation Medicine at UC San Diego Health, a large research and teaching hospital in Western USA, we invited all residents and faculty to participate in our study. In the first step, four faculty and six residents (Table 2) participated via interviews to demonstrate main contouring processes and methods of feedback exchange in residency programs. The same set of faculty and residents also took part in two separate design-thinking workshops that aimed to empower the physicians to reflect on the existing training breakdowns by producing design mock-ups for contouring feedback interfaces. This separation aimed to foster expressing authentic impressions, minimizing the risk of conflict avoidance [85] due to hierarchical power differences between the faculty and residents [48]. The second step involved collecting diverse and granular feedback on the produced mock-ups via a survey study distributed among radiation oncologists globally (including both residents and faculty). The Institutional Review Boards (IRB) approved this study protocol.

3.1 Participants

We recruited participants through the official mailing list of the radiation oncology residency program at the UC San Diego Health, one of the largest research and teaching hospital in the United States. This is one of the largest radiation oncology programs in academic settings, consisting of 12 active faculty and nine residents at the time of the study. To increase traction, one of our collaborators (and an attending faculty in this program) distributed the recruitment call. Four faculty (33.3% acceptance rate) and six residents (66.6%) accepted our invitation to participate in the study.

The residency program at this hospital follows an apprenticeship model of training, in which residents learn contouring and engage in real-world clinical tasks under 1-on-1 supervision from their attending faculty. The residents also rotate with different faculty who are specialized in particular disease sites, such as head/neck and prostate. These rotations can last between 6 weeks and 3 months. Beyond the general structure, the underlying pedagogical and feedback exchange methods are flexible and implemented ad-hoc by the faculty. This paper aims to uncover these methods through interviews from the perspective of faculty and residents.

Besides atlases and guidelines, this residency program lacks specialized learning tools for supporting contouring education. The common tools used by the residents are the same software used for clinical purposes, including MIM and Eclipse as displayed in Figure 2. While these tools provide a plethora of features to assist clinicians in contouring and navigating through medical images, they do not facilitate training and feedback exchange. We especially targeted this gap via the design-thinking workshops in order to explore computer-supported learning interfaces for contouring.

3.2 Study Design

3.2.1 Faculty Interviews. Four radiation oncology faculty participated in one-hour interviews that comprise two steps:

(1) The faculty demonstrated a short contouring session using their preferred software and medical case. They also expressed their thought processes out-loud, such as how they set up contouring sessions, what images they used, and where in the screen they looked. The researchers minimally interrupted, except only when the participants had not spoken for a while, and took notes of key events and explanations. This step familiarized the researchers with the general procedures involved in contouring.

(2) Semi-structured interviews started by asking clarifying questions about the researchers' observations in the

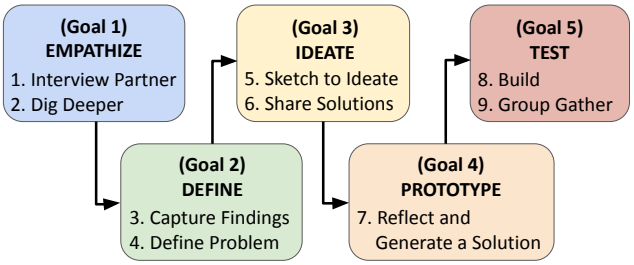


Figure 1: The five goals and nine steps of the design thinking workshops with the radiation oncology faculty and residents. The two workshops aimed to guide participants in designing "ideal contouring feedback interfaces".

think-aloud step. Then, the researchers asked questions that aimed to reveal the faculty's workflow when training residents. Topics included feedback exchange strategies with junior and senior residents, and frequency of training opportunities.

3.2.2 Faculty Workshop. The same four faculty participated in a two-hour remote design thinking workshop that aimed to create their *ideal contouring feedback interfaces*. Design thinking workshops provide a human-centered framework for problem solving [40], and foster exploring needs and ideas for particular stakeholders [51]. Due to the collaborative nature of design thinking methodology, these workshops are commonly conducted in-person, yet circumstances such as pandemics and distant participants can call for remote accommodation.

Inspired by the Wallet project [17], our remote workshop contained five phases (see Figure 1). The two faculty pairs first gained empathy of their partner's contouring practices, and then defined their needs around teaching and learning of contouring skills. After understanding these needs, each pair proceeded to collaboratively generate solution ideas (by sketching designs on Google Doc) and created digital prototypes (using LucidChart[4]). These two tools were shown to be sub-optimal when conducting the remote workshop, as the participants found Google Doc unreliable in terms of formatting, and lacked familiarity with features of LucidChart. Lastly, each pair presented their design to the entire group and received feedback. Overall, two interface mock-ups emerged from this session.

3.2.3 Resident Interviews. Following the faculty workshop, six radiation oncology residents participated in remote, one-hour interviews, following three steps:

(1) The residents first filled out a brief survey on their background information (*e.g.,* age and prior medical school), primary contouring tools, and training strategies (*e.g.,* educational resources and feedback mechanisms).

(2) The residents then contoured a case of their choice without narration while the researchers recorded these sessions. Later, the participants watched these recordings back and provided explanations and thought processes around their contouring decisions and confusion points. Retrospectively thinking out loud aimed to lessen the cognitive load of learners [90], since it can be challenging to simultaneously perform contouring tasks and verbalize thoughts, especially for early residents.

(3) The final 15 minutes prompted resident impressions on the feedback interface mock-ups created during the faculty workshop. The researchers presented and described both prototypes at once, because showing alternative design solutions can produce stronger and more authentic criticisms [86]. The residents then evaluated the two interfaces by describing their desired and undesired features.

3.2.4 Resident Workshop. The design thinking workshop with residents followed the same procedure as the faculty workshop: the researchers introduced the same objective (*i.e.,* designing an ideal contouring feedback interface) and facilitated similar steps (displayed in Figure 1). Due to the logistical challenges faced in the first workshop [96] — *i.e.,* formatting issues and tool unfamiliarity, as described in section 3.2.2) — this workshop incorporated Google Slides for both note-taking and prototyping.

3.2.5 Survey Study. With the goal of enhancing feedback diversity and granularity in the user-centered design protocol of this study, we distributed a survey to collect impressions on the created interface mock-ups. To further elicit *reflective* user feedback — engaging participants beyond surface level "look and feel" concerns [79] — the survey guided the participants through a mix of Likert-scale questionnaires and in-depth tasks of annotation and sketching [87]. The survey was designed using Jotform[5] because of the existing multi-modal features beyond simple text-based questionnaires, and later deployed among 2, 500 most active global users of eContour [77], a popular contouring atlas. The survey contained four sections (and full questions can be found in Appendix B):

(1) *Background information:* The survey started with a demographics section to collect basic background information from survey takers, including age, gender, profession, place of residence, and years of contouring experience.

(2) *Perceived usability and learnability:* The second part of the questionnaire first presented the interfaces and provided short descriptions, and then incorporated four Likert-type scale questions to gauge usability and learnability of each interface, two central pieces in successful design and deployment of learning technology. *Usability* refers to users' evaluation of the usefulness and completeness of interface functions, and *learnability* determines to what extent the respondents preferred the mock-ups for their learning processes. Inspired by surveys on usability [5] and learnability [46], the following questions were incorporated:

- "I think that I would use this interface frequently." (Usability)
- "I found the various functions in this interface well integrated." (Usability)
- "With this interface, I would be more interested to learn the topics." (Learnability)

[4]LucidChart: https://www.lucidchart.com.

[5]JotForm: https://www.jotform.com.

- "With this interface, I would learn to identify the main and important issues of the topic." (Learnability)

While the original questionnaires on usability and learnability contained more questions, incorporating only four statements aimed to reduce the load for the survey takers, which would potentially improve retention and leave more time for the other parts of the survey.

(3) *Liked and disliked features:* To granularly assess perceptions of interface features in each mock-up, the third part of the survey incorporated a brush tool to prompt annotation directly on the interface designs. Two colors were provided: green for "liked" areas, and red for "disliked" regions. Each interface further contained an open-ended text box to enable additional justification on the selected regions.

(4) *Interface design from scratch:* The last section provided space for survey takers to sketch their own "ideal contouring feedback interface", using drawing tools such as free-form pencil, eraser, shapes, and color selector.

3.3 Data Analysis

This section describes the methods used to analyze the qualitative and quantitative data sources.

3.3.1 Interviews. The faculty and residents' interviews contributed to understanding feedback exchange mechanisms in the apprenticeship model of residency training. To examine the semi-structured interviews of the faculty and residents, including residents' perceptions on the mock-ups designed by the faculty, the first author open-coded the transcribed interviews and identified the main topics. Iterative discussions among the team merged these initial codes into preliminary, and then, final themes.

3.3.2 Workshops. To analyze the designed mock-ups that were created as part of the design-thinking workshops, we leveraged two techniques. First, we followed Tohidi *et al.*'s "quick and dirty" [87] method of analysis interface designs, in which we laid out all sketches on a large table, and further re-arranged and grouped designs based on common patterns. Second, we leveraged the final step of the workshops — in which pairs of faculty and residents elaborated on their designs — to draw out underlying reasoning behind the incorporated feedback mechanisms.

3.3.3 Survey. We examined the survey responses according to quantitative and qualitative methods, specifically by running statistical analyses of the Likert-scale questionnaire, creating heatmaps of the annotated regions, and mapping similarity and differences of features across the sketches. The Likert-scale portion of the survey was analyzed using Friedman test [76] — appropriate for ordinal and within-subject data — across the six interface mock-ups per usability and learnability question, followed by pairwise Wilcoxon test [93]. Annotations of liked and disliked regions were filled and overlayed across all responses to create an aggregated depiction of liked and disliked components of each interface mock-up, with green displaying majority liked, red indicating majority disliked, and yellow shades pointing to neutral regions. Meaning, the darker the green, the more positive the evaluation, while the darker the red, the more negative the overall assessment. To granularly examine the final sketches, we identified what features they shared with the original six interface mock-ups which aimed to serve as a metric for functionalities that the physicians found most helpful in their learning of contouring skills (and hence, included in their respective sketches).

4 RESULTS

This section presents results about contouring feedback mechanisms and the overall apprenticeship-based residency training, generated from interviews, the designed mock-up interfaces by the faculty and residents (Figure 3), and reflective-style survey responses. This paper refers to the participants as F1 – F4 for faculty, and R1 – R6 for residents as described in Table 2.

4.1 Three main methods of feedback exchange in residency programs

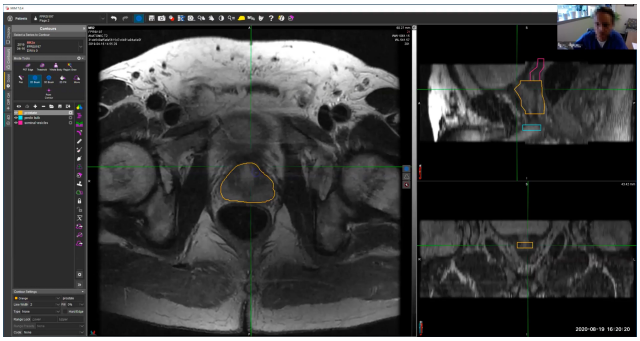
The faculty and resident interviews described three main training strategies as part of the apprenticeship-based model of residency programs, and further unveiled the associated benefits and challenges: 1) assigning clinical cases to residents and later providing contour solutions with additional text-based feedback, 2) contouring sessions where faculty contour and residents watch, and 3) ad-hoc support from senior residents.

Most commonly, the faculty explained that they assigned their own clinical cases to residents, and after residents completed these tasks, the faculty re-contoured the same cases as new structures and sent them along as a source of feedback. F3 explained the benefits of having a visual comparison of both contours for residents: *"they get feedback in terms of looking at what I did versus what they did. [...] I think just over time you sort of develop a skill for looking at these differences and doing the proper windowing"* (F3). He also later emailed his residents to explain the differences, but only if he *"did any major changes"* (F3). While F4 provided similar visual and textual feedback, he emphasized the importance of targeted explanations that reference specific regions in the body:

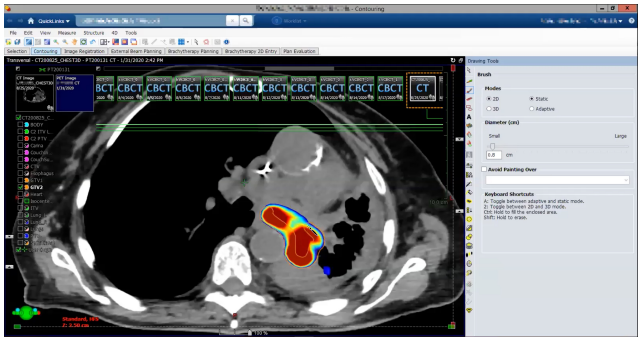
> *"I give specific feedback and since I'm giving them the new structure, even if we're not in person, they can see it. I will say, for example, I deleted the most inferior slice. I don't think that the tumor goes that far. I think that's a vessel."* (F4)

The other two methods of feedback exchange facilitated synchronous faculty-resident and resident-resident interactions. One method involved the resident watching their faculty contour an entire case in a 1-on-1 setting and talk through their strategies, aligned with the *modeling* principle of cognitive apprenticeship (Table 1). Reflecting back on her residency, F2 found this process time-consuming, tiresome, and only marginally beneficial:

> *"As a resident, it's a very tedious and painful process to sit there with your attending and watch them as they adjust pixel by pixel what they want covered and what they don't want covered. And whether or not it's clinically significant, it is up for debate. I used to have an attending that would make me sit with him at the*



(a) A faculty's contouring session on a cancer case using a contouring software called MIM. F4 used a three-image set-up with different orientations of the same set of medical scans.



(b) A resident's contouring session (using Eclipse, another contouring software) on a patient with lung cancer. R4 contoured and viewed a single-image set-up that fused two types of images.

Figure 2: Two examples of the anonymized interview sessions with a faculty (*i.e.,* left image) and a resident (*i.e.,* right image). Both contouring tools contain a main contouring canvas and a number of delineation tools (*e.g.,* brush and eraser) on the side.

> *computer and we'd be there till 10 o'clock at night, and I'd be watching him contour. And to be honest, I don't think I got a lot from it."* (F2)

Lastly, the faculty mentioned that new residents can seek ad-hoc help from more experienced residents whom were more readily available. F1, a new faculty, recalled his early experience as a resident and pointed out the benefits of receiving targeted help (in a back and forth exchange) from the more experienced residents:

> *"The residents all sit in one room. So there's usually two to six residents in the room at any given time. Mostly early on, but less later on, I would grab more senior residents, scroll through images and maybe ask them to help me through one axial variation. Because usually if you're doing one, then it's going to be somewhat similar, meaning once you figure it out for one plane, you can follow it down"* (F1).

R3 also suggested that early assistance can enhance contouring efficiency, especially for new residents: during the contouring phase of the interview, R3 struggled to locate the tumor, and later (in the think-aloud phase) mentioned that *"it's so much easier if you could just ask someone, because I spent too much time trying to find the tumor that might take anyone else like a minute"* (R3).

4.2 Residents favored the visual and descriptive faculty feedback mock-ups

The first faculty pair envisioned an *ideal contouring feedback interface* that aggregates contours (on a single case) and visually maps segments according to the percentage of contours that encapsulated particular regions. As shown in Figure 3a, blue regions represent 20–40% of contours, while the red regions fall within 80–100% of contours. F4 noted the important role of feedback diversity in contouring education: *"it would help residents realize how much variation there is, especially since they only get to work with a handful of attendings"* (F4). The left panel provides further adjustments to the visual representation: different types of interface users (*e.g.,* board-certified users, and second-year residents) can contribute to the distribution map, while contours from specific individuals can overlay the image.

The second faculty pair produced two components in their interface. Figure 3b-left displays a visual comparison between the user contour and the consensus expert contour which highlights the clinical significance of under- and over-contoured areas: exclusion or inclusion of red regions are more problematic than yellow areas. Figure 3b-right provides a text-based description of regions of conflict and their potential long-term impact. It also ranks the user against others with similar levels of experience (*i.e.,* PGY 2, second year residents, in this case). F3 pointed out two unique benefits with this ranking feature, mainly *"drawing on the competitiveness among radiation oncologists or to give you an idea of where you are compared to the other trainees on the same level"* (F3).

Resident interviews revealed that they generally favored both faculty designs, yet weighed the benefits differently with respect to their experience level. More experienced residents identified that the main appeal of Figure 3a was to access a diverse set of perspective on their contours, especially when they only learn from a limited number of faculty:

> *"Typically, the way that residency is structured, you're working one on one with an attending, and so part of it is learning their tendencies, because there's not always one exact right answer. I think that this distribution map is actually a really good idea, because there are those different tendencies and there's not just one right answer, you can see sort of how likely people are to include other structures."* (R4)

Most residents strongly favored Figure 3b mainly due to the emphasis on explaining the contouring differences visually and textually. R4 commented on the shadings for under- and over-contoured regions: *"it is not all about where exactly my contours differ from my attending, but like, why does it matter? Is it an important difference or not?"* (R4). Besides, R2 preferred the text explanations on the right side: *"telling me anatomically, I didn't include the RP lymph nodes or I extended to another part, that is helpful"* (R2). However, some residents raised doubts about the accuracy of the provided long-term impacts, such as R5 (a

third-year resident) who was skeptical about the last statement of the interface: *"if I just saw this, I would be a little skeptical in terms of, where did that come from, how did you decide it is 4% more long term toxicity, as opposed to 8% or 10%"* (R5). Lastly, while residents generally found both designs helpful, they highlighted that each design might satisfy different needs. R3 – who had just started her residency – desired more descriptive feedback:

> *"[Figure 3a] would probably be more useful to someone that's a little bit more advanced in their training versus for me right now, the other one is better, because it gives more information. I just need to know, how I should have done it"* (R3).

4.3 Less experienced residents designed feedback mock-ups to support contouring sessions

The three pairs of residents designed four contouring feedback interfaces (see Figures 3d – 3c). The first pair envisioned a cross-device system that de-couples contouring and feedback. This system contains a *help button* on the top right corner of contouring sessions (Figure 3d-left which shows a work set-up using a large monitor). When uncertainty arises during contouring sessions, residents can press the *help button* and activate feedback on a different device (displayed in Figure 3d-right). This feedback interface determines most similar cases from a medical image database and sorts the images based on similarity to the current case. Two sources populate this image database: cases from resident's attending and general atlases. One member of the pair later elaborated on the significance of highlighting cases of the user's faculty: *"as a resident, you are really only trying to impress your attending"* (R5).

The second pair of residents designed an interface that leverages video for asynchronously capturing more context around residents' questions and experts' answers. This system contains a database of faculty- and resident-created videos. When user faces uncertainty during contouring, they can video record their session: residents can scroll through slices, point to particular regions, and narrate their question. Experts can later go through these video questions and provide answers, either text-based or in video formats (populated under Experts' Videos on the left sidebar in Figure 3f). R2, a member of this pair, justified the video recording feature by emphasizing the benefits of real-time feedback:

> *"While you are contouring a case, all these questions come up, like should I make this adjustment here? should I pull it back anatomically from this structure here? You don't always remember every single question once you are going through it with your attending or you might not have enough time."* (R2)

The third resident pair created two feedback designs: one interface provides tools that support contouring sessions and the other design compares the learner's contour to their attending faculty visually and textually. Figure 3e presents a collection of tools (on the left sidebar) that supports residents during contouring: *Stats* tracks progress and provides hints, *Guidelines* links to relevant external resources, *Similar Cases* presents example prior cases, *Submit* sends the final case for review or radiation planning, and *Share* downloads a de-identified GIF of the case that captures contours on multiple slices. The pair's idea of a de-identified GIF originated from their struggles with software dependency: *"this is just a way to show someone something quickly, so they wouldn't have to be in the hospital and logged into the system"* (R3). The second design (Figure 3c) appears after residents *submit* their contours for feedback: it displays the contours of learner and faculty adjacently and provides description of the differences.

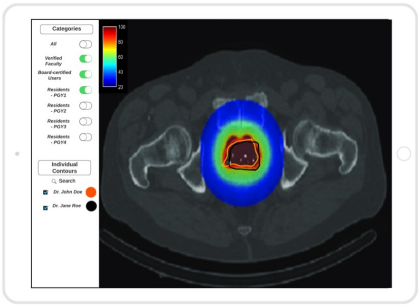
Overall, all four resident mock-ups by large emphasized the importance of *targeted* and *in-session* support, which can differ from the interfaces designed by the experienced faculty that prioritize *aggregated* and *post-hoc* feedback. For instance, Figure 3d (which includes an atlas of similar cases to consult during contouring) and Figure 3f — that facilitates rich multimedia support for capturing and resolving confusions — aim to address contouring breakdowns, especially ones that arise during contouring sessions. The participating faculty, on the other hand, envisioned interfaces that provide feedback post-hoc, once the learner submits their contour for review. These mock-ups especially involved aggregated feedback that captures a wide range of contours, such as the holistic visual representation in Figure 3a, and expert consensus contours and overall prediction of long-term toxicity in Figure 3b.

4.4 Comparison features with similar cases and expert contours can benefit feedback interfaces

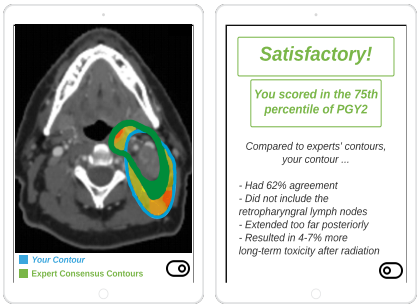
While the interviews helped contextualize the mechanisms of contouring apprenticeship and design-thinking workshops revealed concrete mechanisms to improve feedback exchange, the survey results further shed light on key components of an ideal contouring feedback interface. The survey respondents came from a highly diverse background in terms of gender, profession, years of experience, and especially geographical location. Due to interviews and design-thinking workshops indicating that difference of experience between residents and faculty can affect perception of feedback interfaces, this section considers expertise as a potential factor of analysis.

Demographics — In total, we received 67 survey responses after the survey distribution within the 2,500 most active eContour users (0.029% response rate), comprising 34 female (51%) and 25 male (37%), while 8 responses did not identify a gender. The age range of the respondents were from 22 to 68 years old, with those aged 30 to 40 making up 45% of the total respondents. The participants were primarily practicing radiation oncologists (48; 71%), following 7 medical physicists (11%), and 6 radiation therapists (9%). Contouring experience ranged from 6 months to 28 years, with 27 respondents (40%) having less than or equal to five years of experience.

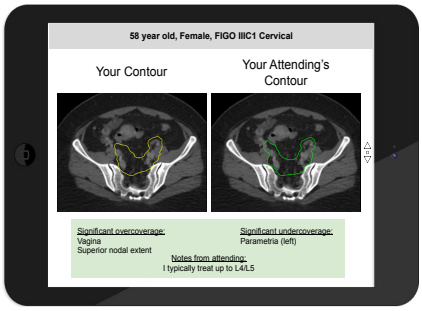
The respondents lived in 31 countries from 6 continents. Europe represented the largest pool with 28 responses (45%) with Russia as the most representative country (8; 13%). The second largest population came from Asia (11; 18%) with India as the most representative country (3; 5%), and North America with the same size of population (11; 18%), including United States (10; 16%) and



(a) Aggregated presentation of contours across all physicians. Multiple user categories and contours from specific individuals can be selected (left). The contouring regions are then color-coded based on distribution of areas covered.

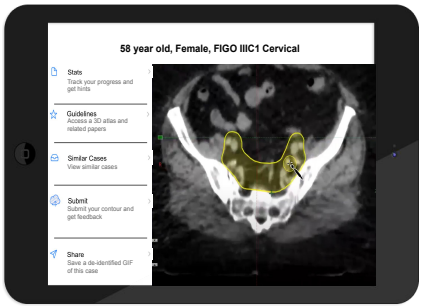


(b) Representation of user and consensus expert contours (left), and textual comparison between the two types (right). Areas of conflict are shaded to highlight clinical significance. The user's contour is ranked against their group.

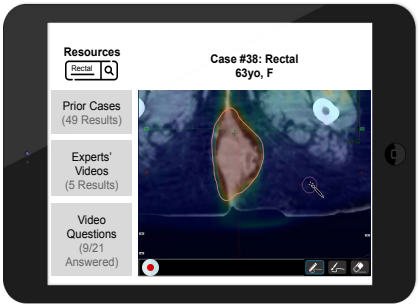


(c) Side-by-side comparison between the contours produced by the user vs. the faculty. In addition to displaying the two contours, it also includes significant contour differences (*i.e.,* over- and under-coverage) and notes from the faculty.

(d) Multi-device interface that displays the most similar cases from the faculty's cases or a general atlas. The help button (top right) activates a second device to sort and present other cases based on similarity.



(e) Tool collection provided to support contouring sessions, such as tracking progress and providing hints, and generating an anonymized motion graphic (*i.e.,* GIF) of the 3D structure.



(f) Video based interface that enables asking questions by recording snippets of contouring sessions, and later receiving feedback via viewing expert videos and consulting similar prior cases.

Figure 3: The generated contouring feedback interfaces in all workshop sessions.

Dominican Republic (1; 2%). The rest of the respondents were from Africa (6; 10%), South America (5; 8%), and Australia (1; 2%).

Perceived Usability and Learnability — We first built an ordinal logistic regression [34] to investigate the effect of two potential independent variables: interface (which is the focus of the Likert-type questions) and expertise, given that the prior workshops pointed to potential differences between how expert faculty and novice residents envision features of an ideal contouring feedback interface. We turned the contouring experience field of the survey questionnaire into three ordered categories, based on the common training model in medical schools: category 1 representing experience level of up to 5 years (i.e., the average length of residency programs), category 2 for 5-10 years of experience to represent the pre-tenured faculty, and category 3 which corresponds to tenured faculty with more than 10 years of contouring experience. Categories 1, 2, and 3 comprised 31 (46.3%), 17 (25.4%), and 19 (28.3%) respondents, respectively. Results show that while *interface* is a predicting factor (b=0.0984, $p < 0.025$), *expertise* does not significantly impact perceived usability and learnability with the p-value of 0.898. Given the significant effect of interface, we then examined how choice of interface impacted each usability and learnability question.

As demonstrated in Figure 4, the Likert-scale questionnaire on the original six interfaces revealed that all interfaces exhibited high levels of usability and learnability. Friedman tests showed significant effect of interface on usability (Q1: $x^2(5) = 26.73, p < 0.001$; Q2: $x^2(5) = 14.15, p < 0.05$), and learnability (Q3: $x^2(5) = 14.07, p < 0.05$; Q4: $x^2(5) = 12.06$, $p < 0.05$). Appendix A displays the pairwise Wilcoxon tests, calculated per question. The results point to interface 5 — containing the resource panel on the left-side as shown in Figure 3e — exhibiting highest levels of usability, given the distributions observed on the figures as well as the significant pairwise differences with the other interfaces. The same interface was also perceived highly in terms of facilitating learning of contouring skills, as displayed in Figure 4. I4 (most similar cases on a separate device, shown in Figure 3d), however, trended towards the lowest perceptions of both usability and learnability.

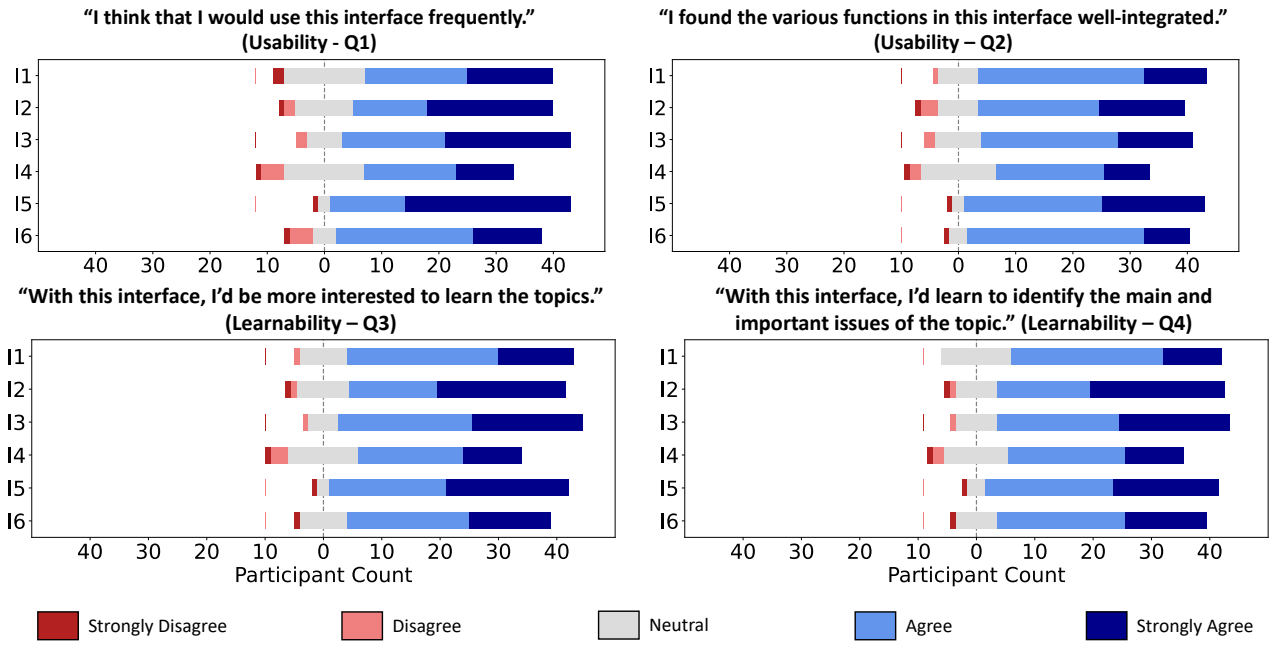


Figure 4: Divergent charts presenting survey answers for the Likert-scale questions regarding *usability* (Q1 & Q2) and *learnability* (Q3 & Q4) of the six mock-ups. Overall differences of distribution are significant, and appendices A1 – A4 present the pairwise tests. I1 – I6 refer to the six designed mockups during the workshops, shown in Figure 5 with added heatmaps collected from the same survey.

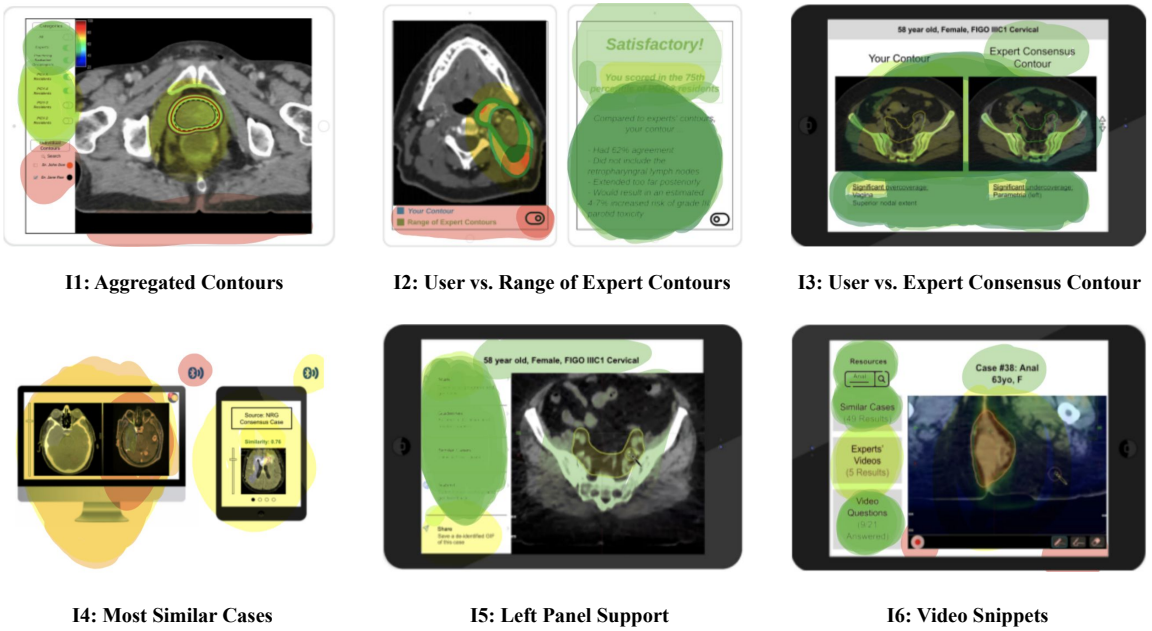
Heatmaps of Liked and Disliked Regions — Filling and overlaying the annotated liked (green) and disliked (red) regions from all responses pointed to granular assessment of the mock-up features. As shown in Figure 5, the participants highly preferred the text-based explanation in interface 2 and 3, meanwhile the multi-device functionality of interface 4 appears to have received a more neutral reaction: one respondent justified disliking this functionality as "separate windows [being] uncomfortable", yet positively rated responses mentioned that it is "interesting to have this [feature]". The participants also favoured aggregating and displaying expert contours (top left in interface 1), and found comparison with the less experienced residents marginally beneficial, as displayed in the middle part of the left panel. Yet, the respondents negatively rated the feature for displaying individual contours (shown in the bottom-left section of interface 1).

Interface Design from Scratch — In total, nine respondents completed the sketching task of the survey which incorporated many design elements from the presented six mock-ups, granularly assessing the benefits of particular functionalities in a contouring feedback interface. As displayed in Figure 6, many sketches pointed to the potential of accessing learning resources *during* contouring sessions and in-situ of the main contouring window, such as guidelines and expert videos in S8 and contouring pearls (*i.e.,* information about case-specific imaging and anatomy) in S9. S4 further developed this principle and envisioned comic-style pop-up hints that spatially reference particular regions on the contouring window and can be toggled on or off. Direct comparison with similar cases was another common theme in many of the sketches (*e.g.,* S2, S5, and S7).

To further analyze the granular components of these sketches, we examined their common features with the original six design mock-ups. All nine sketches incorporated one contouring window as the central piece of the design, similar to I1, I2, I5, and I6. In addition, S1, S2, S6, S8, and S9 showcase the inclusion of the *resource bar* feature observed in I5. Notably, S2 specifically includes *similar cases*, resonating with elements from I4, I5, and I6. S4 and S5 also highlight similarities with the *on/off interactive button* and *select/deselect panel for contour overlays*, respectively, drawing parallels with I1 and I2 features. The *case description* emerges as a focal point in S6, mirroring its prominence in I3, I5, and I6. The *scorecard* — identified as a representative component in I2 — features in S6 and S8 as well. Lastly, S9 introduces the *expert contour* as a distinctive feature, heavily influenced by concepts from I1, I2, I3, and I6.

5 DISCUSSION

This section frames the findings around feedback-exchange tensions in residency programs (Sec. 5.1), dual role of faculty (Sec. 5.2), and cognitive apprenticeship (Sec. 5.3). Specifically, we present how faculty's feedback methods are not in alignment with learners' needs, and later discuss how this misalignment stems from training strategies and content that aim to address clinical duties in addition of teaching, as well as lack of support for learners to examine and share their cognitive processes. Sec. 5.4 describes sociotechnical strategies to improve learning of highly specialized and critical



I1: Aggregated Contours **I2: User vs. Range of Expert Contours** **I3: User vs. Expert Consensus Contour**

I4: Most Similar Cases **I5: Left Panel Support** **I6: Video Snippets**

Figure 5: Heatmap annotations for liked and disliked mock-up regions. The heatmap annotations aimed to directly visualize survey takers' preferences for design elements in the six mock-ups.

medical skills, not just in contouring education, but also in the broader apprenticeship-based healthcare training.

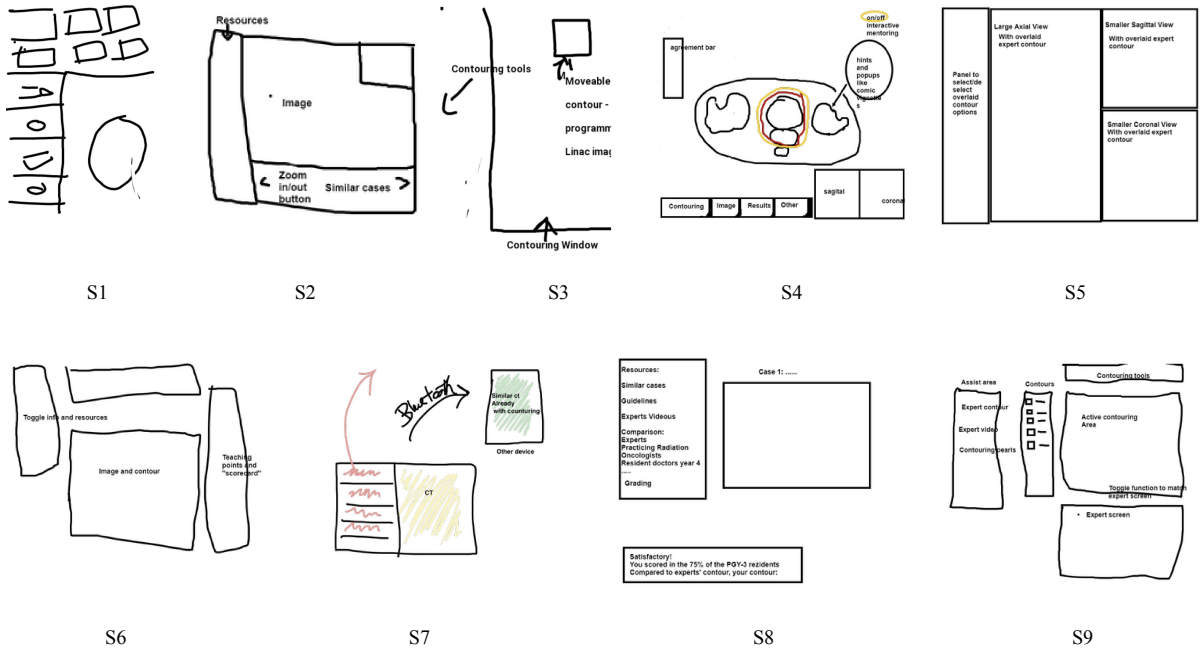
5.1 Tensions between teaching methods of faculty and learning needs of residents

The empirical findings of this work shed light on interrelationships between faculty and residents in the apprenticeship model of residency training, and how the existing mechanisms of feedback exchange do not align with needs of residents, especially in terms of timeliness, relevance, and diversity of training methods.

The asynchronous methods of training introduce significant delay in feedback exchange which leads to subpar in-time support and can degrade overall learning of critical and high-stakes medical tasks. As mentioned in the results section, a common feedback exchange strategy is when faculty assign their own clinical cases to residents as practice opportunities, and later re-contour the entire case and send it back, so the residents can learn by comparing their contour with the expert's. However, this method lacks accounting for confusions and questions that arise when contouring patient cases, evident by interface mock-ups that residents generated during the workshops (*e.g.,* on-demand support features in Figure 3e). Many components of the survey results further showcase the benefit of in-time support, such as high usability and learnability scores of the interface design with the left panel support, as well as the incorporated interactive mentoring functionality that provides hints during contouring sessions (*i.e.,* S4 in Figure 6). Seeking feedback from peer residents is another method of training, yet the ad-hoc nature of this support mechanism can minimize benefits for learners, since there might not be adequate support in place when help is needed, such as unavailability of a senior resident with experience relevant to the case at hand.

The interactions between faculty and residents are limited in supporting unique and granular learning needs. For instance, comparing contours of the entire case with the solution (provided by the faculty) might not directly address gaps of contouring knowledge and skills, since it remains up to the residents to interpret differences as essential concepts or subjective tendencies. The faculty further shared sending notes via email, mainly to provide specific and critical learning points. While the explanation can help clarify some confusions, the barrier to provide detailed and targeted feedback on a disjoint, text-based medium — *i.e.,* email content that needs to map to specific segments of particular images, in a case only accessible by separate contouring tools — can introduce additional burden for the faculty and discourage providing granular feedback. In the training method of watch-alongs, in which faculty think out loud their processes as they place contours on images, these processes can differ from learning needs of the less-experienced residents. While the resident can contribute in this training strategy and ask for clarifications, these questions might differ from the confusions they face when contouring themselves. As such, learning needs can remain unaddressed and only be uncovered when residents engage with contouring tasks in-depth, and when relevant feedback is provided.

Similar to many training programs for specialized healthcare procedures, trainees are matched with only a limited number of expert physicians, which can hinder diversity of feedback especially in complex tasks that involve a certain degree of subjectivity. While access to senior residents can improve variety in feedback, the participants indicated that impressing attending



S1 S2 S3 S4 S5

S6 S7 S8 S9

Figure 6: Nine free-form sketches collected from the survey which shared many elements with the six interface mock-ups.

faculty is a main goal of residency programs. As further unveiled via the workshop (*i.e.,* Figure 3a) and survey (*e.g.,* I1 heatmap in Figure 5), learners valued getting exposed to contouring tendencies and diverse perspectives. In addition, facilitating access to expert physicians (from varying backgrounds and experiences) can substantially improve equity in healthcare: prior research shows that a "quality gap" exists in cancer treatment, in which medical institutions at rural locations (with fewer volume of patients) provide substandard treatment compared to the counterpart urban providers with higher patient volume [1, 50].

5.2 Content and strategies of training that blend pedagogical and clinical duties

The existing misalignment between provided methods of teaching and desired styles of learning can stem from the dual role of clinician and teacher among the attending faculty at medical institutions, as also reported in prior works [66]. These expert physicians are not only expected to provide a quality educational experience to their assigned residents, but also attain a high level of clinical throughput via contouring patient cases, cases that can be particularly critical and time-sensitive. Consequently, when the availability of expert resource is limited, the clinical duties take priority over teaching.

Our findings reveal how the constraint of performing both clinical and pedagogical responsibilities specifically manifests itself by the faculty tailoring the *content* and *strategies* of feedback exchange to also progress through clinical tasks. Evident from the training mechanisms laid out in Sec. 4.1, the use of own clinical cases as educational content, while convenient, might not exactly address the learning needs of residents, given that difficulty, size, and type of case might not be in alignment with the expertise level of residents. Research suggests that educational content that deviates from the medium-difficulty level for learners negatively impacts learning performance [56]. In addition, the feedback strategy of re-contouring the entire case (post residents' submission), while a necessary component of clinical duties, can lack the granularity and depth of feedback that residents need. Contouring watch-alongs can also help satisfy both responsibilities: the faculty can spend time completing clinical tasks, as the resident watches along and marginally benefits in the periphery. As learners elaborated, while thinking out loud about contouring decisions can be helpful, adjusting the contours pixel-by-pixel (on cases that might contain hundreds of slices) takes a long time, time that could be spent on more targeted and specialized practice content.

5.3 Moving from traditional apprenticeships towards cognitive apprenticeship

The findings of this work revealed elements of a residency program that more closely resembles a traditional apprenticeship (*a.k.a.* the first half of a cognitive apprenticeship), in which the three principles of *modeling*, *coaching*, and *scaffolding* are moderately supported. As reported in the results, 1-on-1 contouring watch-alongs — in which experts perform contouring and externalize their internal processes — centers around the first principle (*modeling*). The second step of cognitive apprenticeship model (*coaching*) is partly fulfilled by interacting with peer residents: some participants benefited from working through a small subsection of cases, while the senior resident evaluates their thought process and provides guidance. However, this strategy can be unstructured and ad-hoc, meaning support might not always be available, or expertise of the senior resident can differ from the learner's need. The case exchange and re-contouring method,

while mainly a form of experiential learning (*i.e.,* "learning by doing") [42], shares core elements with the *scaffolding* principle of cognitive apprenticeship, in which the faculty adjust the depth and modality of feedback according to the skill-level of the residents. As noted in the results section, the faculty decide to either only provide the contours, or also add text-based justifications via email exchange when the residents can benefit from the additional hints. Given the asynchronous (*i.e.,* delay in sending feedback) and contextually-limited (*i.e.,* textual *v.s.* richer video formats) nature of the feedback, this training mechanism might not adequately gauge the expertise level of learners in order to provide the relevant help.

The existing training strategies lack support for the second part of cognitive apprenticeship (Table 1), in which learners focus on developing and solidifying their cognitive processes. Given the existing curricular infrastructure, however, residents have limited opportunity to engage in *articulation* and *reflection*, principles that aim to delve deeper into the cognitive processes of learners and enable comparison with experts' model. These steps require investing significant time and resources, an investment that might not directly contribute to clinical throughput, further highlighting the constraint of the dual role of clinician and teacher (as elaborated in Sec. 5.2). Lastly, the *exploration* phase promotes fading, not only in problem-solving, but also in problem-setting, in which learners apply their newly learned skills to seek and tackle other problems that align with their learning goals. This can be difficult, especially in healthcare, for two reasons: first, patient cases involve a high degree of sensitivity and privacy which can pose a barrier for access. Second, it can be challenging to gauge complexity of patient cases, and specifically, what learning goals they cover. As such, residents might need additional guidance in selecting new contouring cases, the type of support that lacks in current training methods. For a complex, critical, and cognitively loaded task like contouring in radiation oncology, as well as clinical workflows in many other healthcare domains, it is imperative that all six principles of cognitive apprenticeship are adequately supported to yield improved learning, and consequently, quality clinical outcomes.

The designed features in the interface mock-ups can especially complement the existing training model by supporting the last three principles of cognitive apprenticeship. For instance, the video-enabled feedback exchange (interface 3f) can be an effective strategy in communicating the internal processes of residents (*i.e., articulation*) and facilitating *reflection* opportunities when enabling learners to compare their cognitive model of expertise (via the proposed feature of Experts' Videos) with their own processes. The Similar Cases feature — introduced in mock-ups 3d and 3e — can also benefit *exploration*, in which residents can explore cases relevant to their current learning task. This approach, however, might require further scaffolding to align these clinical cases with the underlying principles that residents need in improving contouring knowledge and skills.

5.4 Sociotechnical Methods of bridging dual faculty roles and facilitating cognitive apprenticeship

While fundamental solutions for the current residency programs might suggest complete decoupling of clinical and teaching roles among the expert physicians hired at academic institutions, we recognize that these changes require significant re-structuring of existing societal and monetary models and, as such, we offer more attainable curricular and technological strategies to mitigate the shortcomings of healthcare training. This section presents sociotechnical solutions that can, not only address constraints of faculty's dual role (to a reasonable extent), but also support all principles of cognitive apprenticeship. Many of these approaches directly apply to other healthcare domains that incorporate similar apprenticeship models of training.

Leveraging peer resident resource to lower teaching duties and enrich learning — As discovered in the interview sessions, residents seek guidance from their more experienced peers, in a back and forth exchange where the pair can work through a subset of the case together. While especially beneficial for early residents, this method of ad-hoc and informal help-seeking involves additional overhead and uncertainty (*e.g.,* finding senior residents with the relevant level of expertise, when needed) and can further deter residents from pursuing these resources.

Contouring education can especially benefit by systematically leveraging the knowledge and skill-set of senior residents to train a larger number of novice residents. When structured according to case difficulty and expertise level, senior residents can be valuable training resources as they can engage with learners on deeper and longer sessions, and hence, lower the teaching responsibilities of faculty. Many prior HCI works on crowdsourcing explored leveraging the wisdom of expert workers and matching their expertise to the needs of novices by providing concrete learning tasks with representative descriptions, measuring the extent of expert knowledge, and defining reasonable incentives [37, 67, 84]. While residency programs operate at smaller scales than these systems, our findings point to how similar principles can help streamline this process: 1) contouring cases for early residents can be defined by their attending faculty who better understand the complexity of tasks and required skills, 2) senior residents who have specialized in these particular cases can be matched for extended co-contouring sessions that engage learners in deeper cognitive processes, and 3) these expert residents can later be compensated with academic credits or monetary incentives.

Facilitating convenient capture of video snippets to share cognitive processes — As shown during the design-thinking workshops and survey, embedded video recording can help residents capture questions and uncertainties that arise during contouring sessions, and facilitate targeted post-hoc review and learning from the faculty.

Video-assisted feedback is an effective method of feedback exchange in healthcare, as it significantly improves clinical skills [57], and in some cases benefits learners on par with direct expert feedback [64, 65]. In surgery, video summaries — especially

if developed via human leaning models [25] — can benefit resident training by showing alternative ways of performing gestures and enabling residents to trace their mistakes [3]. Creating reusable video snippets, not only lowers the burden on the faculty time in the long-term, but also provides a necessary space for residents' review and self-reflection [24]. These video snippets can especially benefit residency programs, since reviewing and responding via short videos (especially if embedded in the main contouring tools) avoids adding significant overhead to the current workflow of the attending faculty that are responsible for many clinical and teaching tasks.

In addition, convenient video capturing system can facilitate *articulation* principle of a cognitive apprenticeship model, in which learners can express their cognitive processes in-depth and contextually, and further compare their problem-solving skills with experts'. It is particularly beneficial to record these processes in-session, as given the existing training strategies, residents forget many important contouring details and confusions, or there might not be sufficient amount of time during review sessions with the faculty (result presented in Sec. 4.3).

Aggregating variability to capture unique tendencies and yield deliberation — The interviews and interface mock-ups (*e.g.,* the design of 3a and the following positively rated heatmap annotation shown in Figure 5) pointed out the nuanced differences in experts' contours, and how residents raised concerns about the lack of exposure to different contouring tendencies and emphasized learning from diverse styles.

Capturing and presenting other experts' unique contouring tendencies can complement residency programs that facilitate apprenticeship with only a few faculty. Despite existing guidelines (*e.g.,* [49] and [78]), physicians can interpret images differently [44], and hence, introduce contouring variations. As shown in the results, one main source of variation stems from clinicians' dissimilar judgements in including or excluding certain regions around the tumor. Expert disagreements appear in many clinical decision-makings, such as identification of abnormal spikes in brain signals [4] and eye assessment in referral diagnoses [91]. Capturing and presenting contouring disagreements can further encourage deliberation and enhance learning, by especially promoting the *reflection* principle in the cognitive apprenticeship model. Group Deliberation refers to sense-making of the collected uncertainty [29, 72] by leveraging dissenting positions to generate necessary information that can be otherwise lost in consensus-reaching procedures (*e.g.,* majority voting) [80]. In-depth discussions over different contouring tendencies can enable more opportunities for learners to compare their internal cognitive model of expertise with the faculty and peer residents, further aligning with the cognitive apprenticeship model.

An important sociotechnical consideration of collecting variability — according to Ackerman's list of challenges that should be considered in computer supported cooperative work [2] — is critical mass, and specifically in healthcare, scarcity of highly skilled physicians. Critical mass is the idea that a certain threshold of participants is required for the success of a social movement [59] and can affect the perceived usefulness and acceptability of sociotechnical systems [20, 36, 55]. Attracting radiation oncologists, to contribute to a diverse collection of contours, might face challenges due to the lack of critical mass, especially given the already small number of physicians in this field. Careful design of cooperative contouring systems that incorporate elements of the Technology Acceptance Model [16] can enhance user adoption and address critical mass: for instance, since *perceived* critical mass (*e.g.,* through personal interactions) can improve system acceptability [53], feedback solutions (that leverage collection of contours) can start by advertising predominantly to major medical hubs, such as medical schools and oncology clinics.

Providing in-situ and anchored resources to enhance asynchronous faculty feedback — As noted in the interview sessions, a prominent method of feedback exchange is providing solution contours and additional text-based comments provided separately via emails. However, this method can pose learning challenges given the disconnect between the contexts of contouring (via the clinical tools) and the feedback (via email).

The disjoint set of modalities (between the contour solution and textual feedback) can hinder establishing common grounds and exacerbate interlocutors' joint communicative efforts [10, 11]. Prior works in facilitating visual/spatial referencing produced higher quality comments [28], lowered confusion [97], and increased satisfaction [54, 100]. Leveraging the unique characteristics of medical images and spatially anchoring faculty's comments to specific image slices is especially beneficial in contouring residency, in which due to limited availability of the expert faculty with dual roles of clinician and teacher, asynchronous feedback is likely to continue as a prominent training method. An example solution appears in one of the sketches in the survey study (S4 in Figure 6), in which hints are displayed on top of medical images with arrows pointing to specific regions.

Feedback type and presentation can be adjusted to reflect the differing goals of novice and experienced residents. As discussed by Ackerman [2] this is an important consideration for increasing the feasibility of computer supported cooperative systems. Prior research demonstrated how members of organizations can have differing or (sometimes) conflicting goals which can stem from difference of knowledge, meanings, and histories [38, 41, 82]. In contouring education, while targeted and anchored feedback can especially help new residents — who might struggle on region detection and fundamental contouring procedures — experienced learners might benefit more from holistic and diverse feedback. Healthcare training tools should account for the varying goals and experience level of learners to provide effective feedback and avoid disrupting the learning workflow.

6 LIMITATION AND FUTURE WORK

Despite the novel insights that this work extracts and discusses, some sources of limitations exist. All 10 participants (during the interviews and workshops) were from the same medical school, and might have developed similar perceptions about contouring feedback techniques. While the survey study specifically addressed this limitation by engaging clinicians globally, recruiting a larger

and more diverse set of radiation oncologists for in-depth participation can further enhance the external validity of our findings. Future survey studies can also instruct respondents to evaluate interfaces from a particular perspective (*i.e.,* novice *v.s.* expert) given that the level of expertise might impact perception of learning tools. In addition, this paper examined faculty and residents in separate interview and workshop studies. While this allowed us to capture authentic perspectives from both stakeholders given the existing hierarchical power dynamics in residency programs [48], more interactive studies, such as synchronous tutoring simulations, text analysis of email exchanges, and contouring observations might provide deeper insight into the content and techniques of contouring education.

7 CONCLUSION

How is healthcare apprenticeship facilitated in order to transfer highly specialized and critical medical skills, and what are the implications of faculty's dual role of clinician and teacher in mechanisms of feedback exchange? To answer these questions, we examined the inter-dynamics between expert faculty and novice residents in the case of contouring: the high-stakes task of identifying tumours in radiotherapy treatment. Following interviews and design-thinking workshops with faculty (N = four) and residents (N = six), our results revealed tensions between the teaching content and strategies that the faculty provide, and timely, relevant, and diverse support that residents need in order to learn the skills. We describe how this tension arises from overlapping clinical and pedagogical responsibilities of the faculty, and the lack of support for capturing and sharing internal cognitive models of learners. The follow-up survey with practitioners from 31 countries (N = 67) provided diverse perspectives over effective feedback elements of training tools in healthcare.

To resolve the current obstacles, we presented practical sociotechnical solutions that can improve the existing training model in residency programs, including leveraging peer resident resources to lower teaching duties of faculty, facilitating convenient capture of video snippets to share internal cognitive processes of learning, and aggregating variability to yield group deliberation. We believe that understanding the dynamics of apprenticeship training in healthcare is key to improving the quality of training and patient outcome, and future work can especially build on the inherent organizational issues uncovered and discussed in this paper.

REFERENCES

[1] Sahaja Acharya, Samantha Hsieh, Jeff M Michalski, Eric T Shinohara, and Stephanie M Perkins. 2016. Distance to radiation facility and treatment choice in early-stage breast cancer. *International Journal of Radiation Oncology* Biology* Physics* 94, 4 (2016), 691–699.

[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[3] Ignacio Avellino, Sheida Nozari, Geoffroy Canlorbe, and Yvonne Jansen. 2021. Surgical video summarization: multifarious uses, summarization process and ad-hoc coordination. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[4] Elham Bagheri, Justin Dauwels, Brian C Dean, Chad G Waters, M Brandon Westover, and Jonathan J Halford. 2017. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology* 128, 10 (2017), 1994–2005.

[5] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[6] Bennet A Butler, Cameron M Butler, and Terrance D Peabody. 2019. Cognitive apprenticeship in orthopaedic surgery: updating a classic educational model. *Journal of Surgical Education* 76, 4 (2019), 931–935.

[7] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[8] Chen Chen, Matin Yarmand, Varun Singh, Michael V. Sherer, James D. Murphy, Yang Zhang, and Nadir Weibel. 2022. Exploring Needs and Design Opportunities for Virtual Reality-based Contour Delineations of Medical Structures. In *Companion of the 2022 ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Sophia Antipolis, France) *(EICS '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 19–25. https://doi.org/10.1145/3531706.3536456

[9] Chen Chen, Matin Yarmand, Varun Singh, Michael V. Sherer, James D. Murphy, Yang Zhang, and Nadir Weibel. 2022. VRContour: Bringing Contour Delineations of Medical Structures Into Virtual Reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR)* (Singapore) *(ISMAR '22)*. https://doi.org/10.1109/ISMAR55827.2022.00020

[10] Herbert H Clark. 1996. *Using language*. Cambridge university press.

[11] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[12] Allan Collins, John Seely Brown, Ann Holum, et al. 1991. Cognitive apprenticeship: Making thinking visible. *American educator* 15, 3 (1991), 6–11.

[13] Allan Collins, John Seely Brown, and Susan E Newman. 1988. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The journal of philosophy for children* 8, 1 (1988), 2–10.

[14] Allan Collins and Manu Kapur. 2006. *Cognitive apprenticeship*. Vol. 291. na.

[15] Yunfeng Cui, Wenzhou Chen, Lindsey A Olsen, Ronald E Beatty, Peter G Maxim, Timothy Ritter, Jason W Sohn, Jane Higgins, James M Galvin, Ying Xiao, et al. 2015. Contouring variations and the role of atlas in non-small cell lung cancer radiation therapy: Analysis of a multi-institutional preclinical trial planning study. *Practical radiation oncology* 5, 2 (2015), e67–e75.

[16] Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[17] Stanford University Design School. 2016. The wallet project. https://dschool.stanford.edu/resources/the-gift-giving-project

[18] Halil İbrahım Durak, Agah Çertuğ, Ayhan Çalişkan, and Jan Van Dalen. 2006. Basic life support skills training in a first year medical curriculum: six years' experience with two cognitive–constructivist designs. *Medical teacher* 28, 2 (2006), e49–e58.

[19] Pelle Ehn and Morten Kyng. 1987. The collective resource approach to systems design. *Computers and democracy* 17 (1987), 57.

[20] Susan F Ehrlich. 1987. Strategies for encouraging successful adoption of office communication systems. *ACM Transactions on Information Systems (TOIS)* 5, 4 (1987), 340–357.

[21] Marisa E Exter and Iryna Ashby. 2019. Using cognitive apprenticeship to enculturate new students into a qualitative research. *The Qualitative Report* 24, 4 (2019), 873–886.

[22] Alysa Fairchild, Damien C Weber, Raquel Bar-Deroma, Akos Gulyban, Paul A Fenton, Roger Stupp, and Brigitta G Baumert. 2012. Quality assurance in the EORTC 22033–26033/CE5 phase III randomized trial for low grade glioma: the digital individual case review. *Radiotherapy and Oncology* 103, 3 (2012), 287–292.

[23] Jeanne M Farnan, Lindsey A Petty, Emily Georgitis, Shannon Martin, Emily Chiu, Meryl Prochaska, and Vineet M Arora. 2012. A systematic review: the effect of clinical supervision on patient and residency education outcomes. *Academic Medicine* 87, 4 (2012), 428–442.

[24] AL Farquharson, AC Cresswell, JD Beard, and P Chan. 2013. Randomized trial of the effect of video feedback on the acquisition of surgical skills. *Journal of British Surgery* 100, 11 (2013), 1448–1453.

[25] Eléonore Ferrier-Barbut, Ignacio Avellino, Geoffroy Canlorbe, Marie-Aude Vitrani, and Vanda Luengo. 2023. Learning With Pedagogical Models: Videos As Adjuncts to Apprenticeship for Surgical Training. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–40.

[26] Claudio Fiorino, Michele Reni, Angelo Bolognesi, Giovanni Mauro Cattaneo, and Riccardo Calandrino. 1998. Intra-and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning. *Radiotherapy and oncology* 47, 3 (1998), 285–292.

[27] Erin F Gillespie, Neil Panjwani, Daniel W Golden, Jillian Gunther, Tobias R Chapman, Jeffrey V Brower, Robert Kosztyla, Grant Larson, Pushpa Neppala, Vitali Moiseenko, et al. 2017. Multi-institutional randomized trial testing the utility of an interactive three-dimensional contouring atlas among radiation oncology residents. *International Journal of Radiation Oncology* Biology* Physics* 98, 3 (2017), 547–554.

[28] Elena L Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A spatially anchored census of student confusion for online lecture videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1555–1564.

[29] Diego Gracia. 2003. Ethical case deliberation and decision making. *Medicine, Health Care and Philosophy* 6, 3 (2003), 227–233.

[30] Vincent Grégoire, Kian Ang, Wilfried Budach, Cai Grau, Marc Hamoir, Johannes A Langendijk, Anne Lee, Quynh-Thu Le, Philippe Maingon, Chris Nutting, et al. 2014. Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiotherapy and Oncology* 110, 1 (2014), 172–181.

[31] Jeremy M Grimshaw and Ian T Russell. 1993. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *The Lancet* 342, 8883 (1993), 1317–1322.

[32] Jeremy M Grimshaw and Ian T Russell. 1994. Achieving health gain through clinical guidelines II: Ensuring guidelines change medical practice. *Quality in health care* 3, 1 (1994), 45.

[33] Jeremy M Grimshaw, Holger J Schünemann, Jako Burgers, Alvaro A Cruz, John Heffner, Mark Metersky, and Deborah Cook. 2012. Disseminating and implementing guidelines: article 13 in Integrating and coordinating efforts in COPD guideline development. An official ATS/ERS workshop report. *Proceedings of the American Thoracic Society* 9, 5 (2012), 298–303.

[34] UCLA: Statistical Consulting Group. 2021. *ORDINAL LOGISTIC REGRESSION | R DATA ANALYSIS EXAMPLES*. https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/

[35] Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. 85–93.

[36] Jonathan Grudin. 1994. Groupware and social dynamics: Eight challenges for developers. *Commun. ACM* 37, 1 (1994), 92–105.

[37] Philip J Guo. 2015. Codeopticon: Real-time, one-to-many human tutoring for computer programming. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 599–608.

[38] Christian Heath and Paul Luff. 1996. Documents and professional practice: "bad" organisational reasons for "good" clinical records. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. 354–363.

[39] Lisa A Kachnic, Kathryn Winter, Robert J Myerson, Michael D Goodyear, John Willins, Jacqueline Esthappan, Michael G Haddock, Marvin Rotman, Parag J Parikh, Howard Safran, et al. 2013. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. *International Journal of Radiation Oncology* Biology* Physics* 86, 1 (2013), 27–33.

[40] Lucy Kimbell. 2011. Rethinking design thinking: Part I. *Design and culture* 3, 3 (2011), 285–306.

[41] Rob Kling. 1991. Cooperation, coordination and control in computer-supported work. *Commun. ACM* 34, 12 (1991), 83–88.

[42] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

[43] Theodore J Kopcha and Christianna Alger. 2014. Student teacher communication and performance during a clinical experience supported by a technology-enhanced cognitive apprenticeship. *Computers & Education* 72 (2014), 48–58.

[44] Elizabeth A Krupinski. 2000. The importance of perception research in medical imaging. *Radiation medicine* 18, 6 (2000), 329–334.

[45] Chin-Yuan Lai and Yung-Chin Yen. 2018. Using mobile devices to support cognitive apprenticeship in clinical nursing practice–a case study. *Interactive Technology and Smart Education* 15, 4 (2018), 348–362.

[46] Elinda Ai-Lim Lee, Kok Wai Wong, and Chun Che Fung. 2010. How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education* 55, 4 (2010), 1424–1442.

[47] Nancy Y Lee and Jiade J Lu. 2012. *Target volume delineation and field setup: a practical guide for conformal and intensity-modulated radiation therapy*. Springer Science & Business Media.

[48] Heather B Leisy and Meleha Ahmad. 2016. Altering workplace attitudes for resident education (AWARE): discovering solutions for medical resident bullying through literature review. *BMC medical education* 16, 1 (2016), 1–10.

[49] Karen Lim, William Small Jr, Lorraine Portelance, Carien Creutzberg, Ina M Jürgenliemk-Schulz, Arno Mundt, Loren K Mell, Nina Mayr, Akila Viswanathan, Anuja Jhingran, et al. 2011. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. *International Journal of Radiation Oncology* Biology* Physics* 79, 2 (2011), 348–355.

[50] Chun Chieh Lin, Suanna S Bruinooge, M Kelsey Kirkwood, Dawn L Hershman, Ahmedin Jemal, B Ashleigh Guadagnolo, B Yu James, Shane Hopkins, Michael Goldstein, Dean Bajorin, et al. 2016. Association between geographic access to cancer care and receipt of radiation therapy for rectal cancer. *International Journal of Radiation Oncology* Biology* Physics* 94, 4 (2016), 719–728.

[51] Tilmann Lindberg, Christine Noweski, and Christoph Meinel. 2010. Evolving discourses on design thinking: how design cognition inspires meta-disciplinary creative collaboration. *Technoetic Arts: A Journal of Speculative Research* 8, 1 (2010).

[52] Min Liu. 1998. A study of engaging high-school students as multimedia designers in a cognitive apprenticeship-style learning environment. *Computers in Human Behavior* 14, 3 (1998), 387–415.

[53] H Lou, Wenhong Luo, and Diane Strong. 2000. Perceived critical mass effect on groupware acceptance. *European journal of information systems* 9, 2 (2000), 91–103.

[54] Kurt Luther, Amy Pavel, Wei Wu, Jari-lee Tolentino, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2014. CrowdCrit: crowdsourcing and aggregating visual design critique. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. 21–24.

[55] M Lynne Markus and Terry Connolly. 1990. Why CSCW applications fail: Problems in the adoption of interdependent work tools. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*. 371–380.

[56] Janet Metcalfe and Nate Kornell. 2003. The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General* 132, 4 (2003), 530.

[57] Craig I Nesbitt, Alexander W Phillips, Roger F Searle, and Gerard Stansby. 2015. Randomized trial to assess the effect of supervised and unsupervised video feedback on teaching practical skills. *Journal of surgical education* 72, 4 (2015), 697–703.

[58] CF Njeh. 2008. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of medical physics/Association of Medical Physicists of India* 33, 4 (2008), 136.

[59] Pamela Oliver, Gerald Marwell, and Ruy Teixeira. 1985. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American journal of Sociology* 91, 3 (1985), 522–556.

[60] Annemarie Sullivan Palincsar. 1986. Metacognitive strategy instruction. *Exceptional children* 53, 2 (1986), 118–124.

[61] Aannemarie Sullivan Palinscar and Ann L Brown. 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and instruction* 1, 2 (1984), 117–175.

[62] N. Panjwani, E.F. Gillespie, D.W. Golden, J.R. Gunther, T.R. Chapman, J.V. Brower, J.M. Bykowski, P. Sanghvi, and J.D. Murphy. 2016. Usability of a Novel Interactive Web-Based Contouring Atlas. *Internation Journal of Radiation Oncology, Biology and Physics* (2016). https://doi.org/10.1016/j.ijrobp.2016.06.1675

[63] Lester J Peters, Brian O'Sullivan, Jordi Giralt, Thomas J Fitzgerald, Andy Trotti, Jacques Bernier, Jean Bourhis, Kally Yuen, Richard Fisher, and Danny Rischin. 2010. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *Journal of clinical oncology* 28, 18 (2010), 2996–3001.

[64] Alexander W Phillips, Joanna Matthan, Lucy R Bookless, Ian J Whitehead, Anantha Madhavan, Paul Rodham, Anna LR Porter, Craig I Nesbitt, and Gerard Stansby. 2017. Individualised expert feedback is not essential for improving basic clinical skills performance in novice learners: a randomized trial. *Journal of surgical education* 74, 4 (2017), 612–620.

[65] James Rammell, Joanna Matthan, Matthew Gray, Lucy R Bookless, Craig I Nesbitt, Paul Rodham, John Moss, Gerard Stansby, and Alexander W Phillips. 2018. Asynchronous unsupervised video-enhanced feedback as effective as direct expert feedback in the long-term retention of practical clinical skills: randomised trial comparing 2 feedback methods in a cohort of novice medical students. *Journal of surgical education* 75, 6 (2018), 1463–1470.

[66] Kate Rassie. 2017. The apprenticeship model of clinical medical education: time for structural change. *The New Zealand Medical Journal (Online)* 130, 1461 (2017), 66.

[67] Daniel Rees Lewis, Emily Harburg, Elizabeth Gerber, and Matthew Easterday. 2015. Building support tools to connect novice designers with professional coaches. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 43–52.

[68] Wolff-Michael Roth and G Michael Bowen. 1995. Knowing and interacting: A study of culture, practices, and resources in a grade 8 open-inquiry science classroom guided by a cognitive apprenticeship metaphor. *Cognition and instruction* 13, 1 (1995), 73–128.

[69] Kittima Sadhuwong, Prakob Koraneekij, and Onjaree Natakuatoong. 2016. Effects of a blended learning model integrating situated multimedia lessons and cognitive apprenticeship method on the clinical reasoning skills of nursing students. *Journal of Health Research* 30, 6 (2016), 421–431.

[70] Marlene Scardamalia. 1987. *The psychology of written composition*. Hillsdale, NJ: L. Erlbaum Associates.

[71] Marlene Scardamalia, Carl Bereiter, and Rosanne Steinbach. 1984. Teachability of reflective processes in written composition. *Cognitive science* 8, 2 (1984), 173–190.

[72] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding expert disagreement in medical data

analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[73] Joseph R Schneider, John J Coyle, Elizabeth R Ryan, Richard H Bell Jr, and Debra A DaRosa. 2007. Implementation and evaluation of a new surgical residency model. *Journal of the American College of Surgeons* 205, 3 (2007), 393–404.

[74] Alan H Schoenfeld. 1983. Problem Solving in the Mathematics Curriculum. A Report, Recommendations, and an Annotated Bibliography. MAA Notes, Number 1. (1983).

[75] Alan H Schoenfeld. 2016. Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (Reprint). *Journal of education* 196, 2 (2016), 1–38.

[76] Michael R Sheldon, Michael J Fillyaw, and W Douglas Thompson. 1996. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International* 1, 4 (1996), 221–228.

[77] Michael V Sherer, Diana Lin, Kartikeya Puri, Neil Panjwani, Zhigang Zhang, James D Murphy, and Erin F Gillespie. 2019. Development and usage of eContour, a novel, three-dimensional, image-based web site to facilitate access to contouring guidelines at the point of care. *JCO Clinical Cancer Informatics* 3 (2019), 1–9.

[78] William Small Jr, Loren K Mell, Penny Anderson, Carien Creutzberg, Jennifer De Los Santos, David Gaffney, Anuja Jhingran, Lorraine Portelance, Tracey Schefter, Revathy Iyer, et al. 2008. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy in postoperative treatment of endometrial and cervical cancer. *International Journal of Radiation Oncology* Biology* Physics* 71, 2 (2008), 428–434.

[79] Hilary Smith, Geraldine Fitzpatrick, and Yvonne Rogers. 2004. Eliciting reactive and reflective feedback for a social communication tool: a multi-session approach. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*. 39–48.

[80] Miriam Solomon. 2006. Groupthink versus the wisdom of crowds: The social epistemology of deliberation and dissent. *The Southern journal of philosophy* 44, S1 (2006), 28–42.

[81] Renée E Stalmeijer. 2015. When I say… cognitive apprenticeship. *Medical Education* 49, 4 (2015), 355–356.

[82] Susan Leigh Star and Karen Ruhleder. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research* 7, 1 (1996), 111–134.

[83] Myat Su Yin, Peter Haddawy, Siriwan Suebnukarn, and Phattanapon Rhienmora. 2016. Toward intelligent tutorial feedback in surgical simulation: Robust outcome scoring for endodontic surgery. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 402–406.

[84] Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. 2016. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2645–2656.

[85] Devin T Sydor, Viren Naik, and Zeev Friedman. 2015. Residents' reluctance to challenge negative hierarchy in the operating room: a qualitative study. *Canadian Journal of Anesthesia* 62, 6 (2015), 576.

[86] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1243–1252.

[87] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. User sketches: a quick, inexpensive, and effective way to elicit more reflective user feedback. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*. 105–114.

[88] Andrew T Trout, Page I Wang, Richard H Cohan, Janet E Bailey, Shokoufeh Khalatbari, Jamie D Myles, and N Reed Dunnick. 2011. Apprenticeships ease the transition to independent call: an evaluation of anxiety and confidence among junior radiology residents. *Academic Radiology* 18, 9 (2011), 1186–1194.

[89] Chun-Yen Tsai, Brady Michael Jack, Tai-Chu Huang, and Jin-Tan Yang. 2012. Using the cognitive apprenticeship web-based argumentation system to improve argumentation instruction. *Journal of Science Education and Technology* 21 (2012), 476–486.

[90] Maaike Van Den Haak, Menno De Jong, and Peter Jan Schellens. 2003. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology* 22, 5 (2003), 339–351.

[91] Monica Van Such, Robert Lohr, Thomas Beckman, and James M Naessens. 2017. Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice* 23, 4 (2017), 870–874.

[92] John Whiteside, John Bennett, and Karen Holtzblatt. 1988. Usability engineering: Our experience and evolution. In *Handbook of human-computer interaction*. Elsevier, 791–817.

[93] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[94] Evan J Wuthrick, Qiang Zhang, Mitchell Machtay, David I Rosenthal, Phuc Felix Nguyen-Tan, André Fortin, Craig L Silverman, Adam Raben, Harold E Kim, Eric M Horwitz, et al. 2015. Institutional clinical trial accrual volume and survival of patients with head and neck cancer. *Journal of Clinical Oncology* 33, 2 (2015), 156.

[95] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[96] Matin Yarmand, Chen Chen, Danilo Gasques, James D Murphy, and Nadir Weibel. 2021. Facilitating remote design thinking workshops in healthcare: the case of contouring in radiation oncology. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[97] Matin Yarmand, Srishti Palani, and Scott Klemmer. 2021. Adjacent Display of Relevant Discussion Helps Resolve Confusion. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–11.

[98] Matin Yarmand, Michael Sherer, Chen Chen, Larry Hernandez, Nadir Weibel, and James D. Murphy. 2022. Evaluating Accuracy, Completion Time and Usability of Everyday Touch Devices for Contouring. *International Journal of Radiation Oncology, Biology, Physics* 114, 3 (2022), S96.

[99] Matin Yarmand, Borui Wang, Chen Chen, Michael Sherer, Larry Hernandez, James Murphy, and Nadir Weibel. 2023. Design and Development of a Training and Immediate Feedback Tool to Support Healthcare Apprenticeship. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[100] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S Fels. 2019. " Can you believe [1: 21]?!" Content and Time-Based Reference Patterns in Video Comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

A PAIRWISE WILCOXON TESTS FOR LIKERT-SCALE QUESTIONS IN THE SURVEY

This section presents the pairwise Wilcoxon tests for the the four Likert-scale questions in the survey:

Question 1: *"I think that I would use this interface frequently."* (see Table A1).

Question 2: *"I found the various functions in this interface well integrated."* (see Table A2)

Question 3: *"With this interface, I would be more interested to learn the topics."* (see Table A3)

Question 4: *"With this interface, I would learn to identify the main and important issues of the topic."* (see Table A4)

B SURVEY INSTRUCTIONS

B.1 Step 1 of 3: Demographics

Please provide the following background information. This helps us contextualize your responses later in the survey.

- How old are you?
- What is your gender?
- What is your affiliated industry/academic institution?
- What is your job title?
- How long have you been contouring?

B.2 Step 2 of 3 (a): Demographics

Here, you will see six contouring feedback interfaces. Each image contains a description on the right side of the image. Please familiarize yourself with these designs before moving on to the next questions.

B.3 Step 2 of 3 (b): Perceived Usability and Learnability

Please answer the following prompts by navigating the drop-down menu on each interface.

- I think that I would use this interface frequently.
- I found the various functions in this interface well integrated.
- With this interface, I would be more interested to learn the topics.
- With this interface, I would learn to identify the main and important issues of the topic.

B.4 Step 2 of 3 (c): Interface Annotations

In this section, please evaluate specific components of the 6 interfaces above. For each design:

- Use the pencil tool to specify what parts of the interface you like and dislike. You can draw around the components of your choice with the colours green (for regions that you like) and red (for regions that you dislike).
- Explain your reasoning for the liked and disliked regions underneath the images. The left column is for liked regions and the right column is for disliked regions.

B.5 Step 3 of 3: Sketching

By this final stage of the survey, you have seen six feedback interfaces. It is now your turn! Use the space provided to design YOUR ideal contouring feedback interface.
Don't worry about creating a professional-looking design! A quick sketch/drawing that illustrates the essential elements of your interface would be sufficient. You can even choose to insert text boxes in place of complex drawing components.

Table A1: Survey results of "I think that I would use this interface frequently." (Question 1). We used "*", "", and "***" to indicate statistical significance of $p < .05$, $.01 < p < .05$, and $p < .001$.**

	I1	I2	I3	I4	I5	I6
I1	1.0	0.22	0.089	0.49	**0.000045 (***)**	0.65
I2	0.22	1.0	0.63	0.056	**0.0037 (**)**	0.44
I3	0.089	0.63	1.0	**0.017 (*)**	**0.015 (*)**	0.21
I4	0.49	0.056	**0.017 (*)**	1.0	**0.0000020 (***)**	0.25
I5	**0.000045 (***)**	**0.0037 (**)**	**0.015 (*)**	**0.0000020 (***)**	1.0	**0.00027 (***)**
I6	0.65	0.44	0.21	0.25	**0.00027 (***)**	1.0

Table A2: Survey results of "I found the various functions in this interface well integrated." (Question 2). We used "*", "", and "***" to indicate statistical significance of $p < .05$, $.01 < p < .05$, and $p < .001$.**

	I1	I2	I3	I4	I5	I6
I1	1.0	0.99	0.77	0.10	0.059	0.90
I2	0.99	1.0	0.78	0.11	0.058	0.89
I3	0.77	0.78	1.0	0.18	**0.030 (*)**	0.68
I4	0.10	0.11	0.18	1.0	**0.00050 (***)**	0.08
I5	0.059	0.058	**0.030 (*)**	**0.00050 (***)**	1.0	0.08
I6	0.90	0.89	0.68	0.08	0.08	1.0

Table A3: Survey results of "With this interface, I would be more interested to learn the topics." (Question 3). We used "*", "", and "***" to indicate statistical significance of $p < .05$, $.01 < p < .05$, and $p < .001$.**

	I1	I2	I3	I4	I5	I6
I1	1.0	0.23	0.21	0.36	**0.013 (*)**	0.49
I2	0.23	1.0	0.95	**0.037 (*)**	0.19	0.61
I3	0.21	0.95	1.0	**0.032 (*)**	0.21	0.57
I4	0.36	**0.037 (*)**	**0.032 (*)**	1.0	**0.00075 (***)**	0.11
I5	**0.013 (*)**	0.19	0.21	**0.00075 (***)**	1.0	0.070
I6	0.49	0.61	0.57	0.11	0.070	1.0

Table A4: Survey results of "With this interface, I would learn to identify the main and important issues of the topic." (Question 4). We used "*", "", and "***" to indicate statistical significance of $p < .05$, $.01 < p < .05$, and $p < .001$.**

	I1	I2	I3	I4	I5	I6
I1	1.0	**0.025 (*)**	0.11	0.99	**0.014 (*)**	0.16
I2	**0.025 (*)**	1.0	0.50	**0.024 (*)**	0.83	0.40
I3	0.11	0.50	1.0	0.11	0.38	0.87
I4	0.99	**0.024 (*)**	0.11	1.0	**0.014 (*)**	0.15
I5	**0.014 (*)**	0.83	0.38	**0.014 (*)**	1.0	0.29
I6	0.16	0.40	0.87	0.15	0.29	1.0