AN ABSTRACT OF THE DISSERTATION OF

Anahita Sanandaji for the degree of Doctor of Philosophy in Computer Science presented on June 8, 2018.

Title: Developing a 2D Cross-section Training Strategy for 3D Volume Segmentation by Analyzing Human Perception and Cognitive Tasks

Abstract approved: _____

Cindy Grimm

3D volume segmentation is a fundamental process in many scientific and medical applications. Producing accurate segmentations, in an efficient way, is challenging, in part due to low imaging data quality (e.g., noise and low image resolution), and ambiguity in the data that can only be resolved with higher-level knowledge of the structure. Automatic algorithms do exist, but there are many use cases where they fail. The gold standard is still manual segmentation or review. Unfortunately, even for an expert, manual segmentation is laborious, time consuming, and prone to errors. Existing 3D segmentation tools are often designed based on the underlying algorithm, and do not take into account human mental models, their lower-level perception abilities, and higher-level cognitive tasks.

In this research, we analyzed manual segmentation as a human-computer interaction paradigm to gain a better understanding of both low-level (perceptual) actions, and higher-level tasks and decision-making processes. We initially employed formative field studies using our novel hybrid protocol that blends observation, surveys, and eye-tracking. We then developed, and validated, data coding schemes to discern segmenters' low-level actions, higher-level tasks, and overall task structures. Using these methods, we successfully identified different segmentation strategies utilized by the segmenters. In addition, formative study results showed that the ability to understand 2D cross-sections of 3D structures is a necessary skill in 3D volume segmentation that can be improved through practice and training.

We used the results of our formative studies to introduce a domain-agnostic 2D cross-section training strategy for 3D volume segmentation and developed an interactive training tool to help novices correctly identify 2D cross-sections of 3D structures.

To evaluate the effectiveness of our training tool, we designed a novel 2D cross-section test instrument based on various spatial ability factors. We then conducted user studies and used the test instrument to measure participants' performance before and after the training. Study results show that the training tool is effective in improving participants' 2D cross-section understanding skills, which then can be used to perform a more accurate 3D volume segmentation.

Developing a 2D Cross-section Training Strategy for 3D Volume Segmentation by Analyzing Human Perception and Cognitive Tasks

by

Anahita Sanandaji

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 8, 2018
Commencement June 2019

Doctor of Philosophy dissertation of Anahita Sanandaji presented on June 8, 2018.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Anahita Sanandaji, Author

ACKNOWLEDGEMENTS

First of all, I would like to thank my academic adviser, Dr. Cindy Grimm, for all her help, guidance, and support throughout my entire journey at Oregon State University. Her mentorship and encouragement have helped me grow into a more capable researcher. What I learned from her is far from just performing scientific research and will stay with me for my life.

Thanks to my committee, Dr. Carlos Jensen, Dr. Eugene Zhang, and Dr. Ronald Metoyer for their support, insightful comments, and encouragement. Many thanks goes to Dr. Ruth West, professor of University of North Texas (UNT), for her consistent help and suggestions on conducting the studies and analyses, writing the papers, and pushing this research forward. A special thanks goes to Dr. Christopher Sanchez for providing many insights into this research. I would also like to thank the UNT undergraduate team for helping me run the studies and analyze the data.

Finally, it would take more than an acknowledgement to thank my family and friends for their endless love and support. My heartfelt thanks to Nourieh, for her hospitality and unconditional love while I was in Corvallis. Thanks to my Dad, Farzad, for raising me with the desire to seek knowledge and wisdom. Thanks to my Mom, Mahboobeh, for the sacrifices she made to improve her children's educations. Thanks to my brother, Nader, and my sister, Mandana, who have never left my side and their presence has filled my heart with joy. I am also very blessed to have my life teammate, Saeed, who stands with me and brightens my life, to share this accomplishment with.

This research was funded by NSF Grants IIS 1302142 and IIS 1302248.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF ALGORITHMS

Algorithm Page

Chapter 1: Introduction

Recent advances in imaging techniques have made a fundamental impact on both scientific research and clinical practice. Imaging produces spatial data which, in its raw form, only consists of colored pixels or voxels. Therefore, the information from imaging needs to be processed to reveal useful knowledge about the imaged subject. The process starts with the delineation of the anatomical structure of interest (e.g., an organ, a tissue, a cell, or a collection of these). This is known as 3D volume segmentation.

3D volume segmentation is not only a fundamental step in biomedical imaging (e.g. locating tumors and other pathologies, measuring tissue volumes and computer guided surgery), but is also critical for a wide range of qualitative and quantitative scientific applications including: 1) Visualization: Enabling accurate rendering of the segmented structures with a variety of styles (e.g., show or hide, transparent or surface only); 2) Quantitative analysis: Gathering quantitative information about the structure, including: geometry (e.g., orientation, surface area, and interior volume), topology (e.g., connectivity and branching), and high-level metrics (e.g., shape and morphology); and 3) Physical computation: Using segmentations for treatment plan optimizations and virtual simulations (e.g., fluid dynamics and physical deformations) that explore the physical properties of the anatomical structures and how they interact with the environment.

Figure 1.1 shows examples of real-world segmentation problems with varying quality of structure boundaries, geometric complexity, and number of structures to be segmented. For many medical applications, it is crucial to obtain an accurate segmentation in an efficient way or otherwise health risks in clinical practices increases [30, 31]. Radiation therapy in cancer treatment is a good example that shows the importance of accuracy and efficiency in a segmentation process. From a patient's CT scan, radiologists perform segmentation to build 3D models of the tumors as well as normal organs and tissues. These models are then used to carefully optimize direction and dose of radiation to attack tumors while avoiding healthy tissues. The segmentation process therefore needs to be accurate. The entire process should also be as short as possible to minimize the chance of changes in shape and size of tumors between when the patient is scanned and when

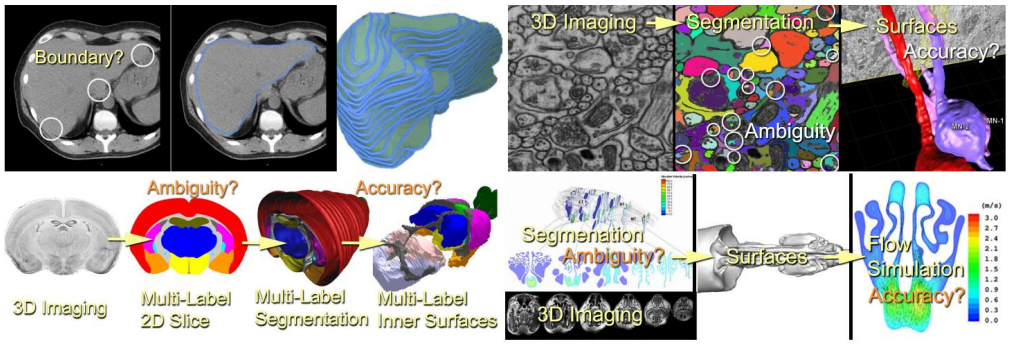


Figure 1.1: Examples of real-world segmentation. Top left: Liver in a CT volume. Top Right: Neurons (electron tomography). Bottom left: Multiple anatomical regions in a mouse brain from sections. Bottom right: Rabbit nasal passage airway in a NMR volume. Flow is visualized on each slice.

the radiation treatment is performed.

It is a non-trivial task to create accurate 3D segmentations of biomedical data in an efficient way. The segmentation process might involve complex or unclear structures (e.g., the boundaries of biological tissues are not always clearly separated as shown in Figure 1.2).

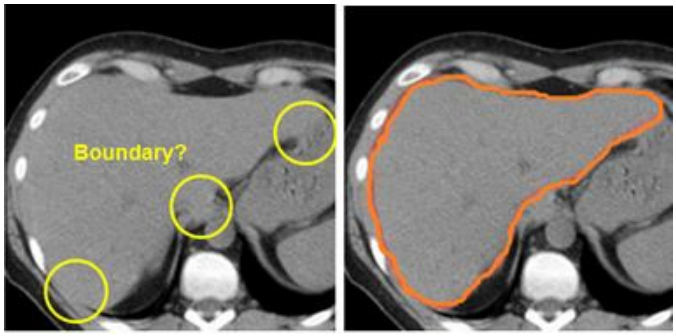


Figure 1.2: Liver data set (right) and its contour (left). The boundaries are not clearly separated in the right image.

Low image resolution and noise also introduce serious issues with the segmentation process. In addition, the segmentation process involves working with 2D slices of 3D data and it is difficult for humans to mentally integrate them into a coherent 3D structure. While humans are adept at visual recognition in the real world (e.g. identifying faces

and objects in familiar scenes), identifying complex anatomical structures in biomedical images is still challenging for them (see Figure 1.3). As a result, 3D segmentation of biomedical data is usually done by experts who have many years of experience in segmenting specific anatomical structures.

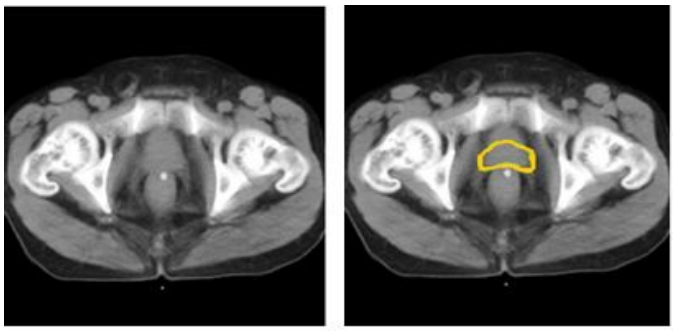


Figure 1.3: Example of a complex anatomical structure. Left: CT scan of a male pelvis. Right: Yellow contour shows the location of prostate. The contour has been drawn by an expert. A novice cannot identify the location of prostate without having previous knowledge of the structure.

Human experts can perform this task using complex analysis of shape, intensity, position, texture, and proximity to surrounding structures [107]. They create 3D segmentations using low-level perceptual cues to perform low-level marking, all guided by higher-level cognitive tasks and domain knowledge. Low-level cues include texture or spatial attributes [87], low-level marking tasks include delineating structures in an image plane by marking contours or filling regions. Experts also use higher-level constraints such as connectivity, shape and topology to disambiguate unclear structure boundaries [48]. So, one approach for manual segmentation is to manually draw contours on selected cross-sections of a 3D volume. Figure 1.4 summarizes the process [108]. Even with experts, segmentation is a time-consuming process, as they usually explicitly mark the structure boundary throughout the entire volume.

1.1 Research Motivation

Despite the large amount of work on developing automated and interactive segmentation tools (see Section 1.2), there is little research on what is the correct way to design, imple-

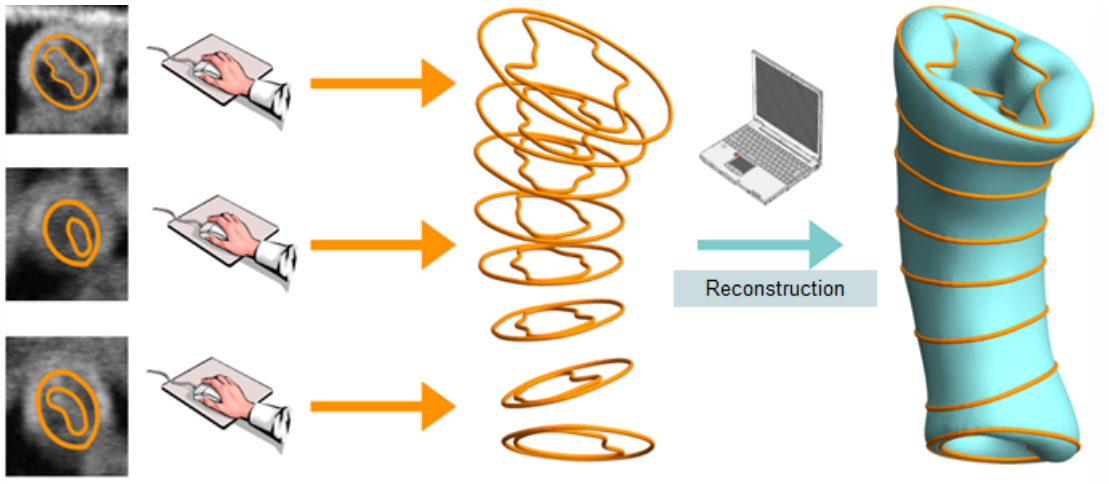


Figure 1.4: Manual Segmentation process. An expert draws contours on selected 2D cross-sections of a 3D volume. Then, using a reconstruction algorithm, the 3D volume of the structure is created [108].

ment, and evaluate, such tools. There is enough evidence that the current segmentation process is both inefficient [62], and lacks repeatability [13, 31, 45, 98, 101]. The causes of variability, inefficiencies and errors in a segmentation process are not fully examined. In addition, the current design and evaluation methods do not capture all aspects involved in 3D volume segmentation process, including human factors. The role of humans in 3D segmentation is undeniable. After segmenting an image either manually or automatically, an expert reviews and inspects [92] the result and decides whether it is acceptable or not. Then, if necessary, he/she modifies and corrects the results (e.g. by using the graphical editor of the segmentation tool [65]). Unfortunately, 3D segmentation design and evaluation methods pay little attention to the role of humans, their mental model and how they perform low-level tasks and define the higher level criteria while doing 3D segmentation. This lack of understanding hinders segmentation tools' development and improvement.

The purpose of our research is to investigate 3D volume segmentation as a human-computer interaction paradigm and to understand human factors that are involved in the current segmentation process, with the goal of making the process more efficient, accurate and repeatable. It is important to study humans independently of a single

segmentation tool, to better understand the process. We argue that segmentation and training process could be significantly improved with a better understanding of how human segmenters perform low-level perception and higher level cognition tasks in the context of 3D segmentation (e.g., visual cues, delineation of structures by marks, 2D cross-section, and 3D structure understanding), and use their mental model to guide local decision makings for a better segmentation.

1.2 Background

In this section we first present related literature on different segmentation methods and tools. Then, we review 3D image segmentation evaluation criteria and approaches. Finally, we talk about human factors in 3D volume segmentation.

1.2.1 3D Volume Segmentation Methods and Tools

There are many different proprietary and open source tools and software for image segmentation. Examples of proprietary tools are MIPAV (Medical Image Processing, Analysis, and Visualization) [1], Amira [91] and TurtleSeg [72]. Some open source tools for image segmentations are ITK (Insight Segmentation and Registration Toolkit) [47], VTK (Visualization Toolkit) [83], ITK-SNAP [105], 3D Slicer [25], and VolumeViewer [89].

Manual segmentation is still a gold standard for image segmentation (specifically for medial imaging). However, as segmenters typically mark the structure boundary throughout the entire image, the manual segmentation is a time consuming process. Automatic segmentation algorithms and methods can improve the efficiency of manual segmentation [2, 15, 32, 103]. There exists various methods for automatic segmentation, including thresholding, edge tracking, region growing, k-means clustering, deformable models, watershed segmentation, graph cut algorithms, shape models, appearance models and atlas-based segmentation [69]. Unfortunately, automated solutions do not provide the same level of accuracy as humans in the presence of image noise and poorly defined boundaries.

Due to the limitation of automatic methods, several interactive or semi-automated approaches have been developed to benefit from the strength of both human and computers. These methods allow the segmenter to specify constraint points by sketching

foreground and background regions on a cross-sectional plane or the volume rendered image or by sketching a contour of the region of interest on the volume [90]. For example, in an interactive segmentation method, the expert may draw markers on selected 2D slices of the original 3D image, to delineate the interior or boundary of the structure. Then, an algorithm is employed to do semi-automated segmentation based on markers. After that, the expert checks the result of the segmentation, does necessary modifications, and continues using the algorithm till reaching a satisfactory result.

Many scientific sites and clinical labs also have their own in-house semi-automated segmentation tools. These in-house tools are designed and optimized in a way to cover specific segmentation needs of the site including different structural scales, imaging modalities, and end-use applications. However, there is no global mechanism to evaluate these tools.

1.2.2 3D Image Segmentation Evaluation

To understand the quality and user variability for 3D image segmentation, researchers have developed a number of common evaluation criteria such as: Accuracy, repeatability and efficiency [65].

Accuracy: The most common criterion for evaluating the results of segmentation process is accuracy or validity. Accuracy is a degree to which the result of segmentation matches the truth [90]. Olabarriaga and Smeulders [65], mentioned that accuracy can be assessed subjectively or objectively, depending on whether a human expert checks the segmented image visually or not. Figure 1.5 shows a more comprehensive hierarchy of evaluating accuracy introduced in [106].

Repeatability: Also known as consistency or precision, repeatability is the extent to which the same result is produced from different segmentation processes performed by the same or different users [90]. We can measure consistency in terms of intra-observer and inter-observer variability.

Efficiency: Also called viability [95], efficiency is the extent to which time and effort are well used during the segmentation process. The total elapsed time is an indicator of overall efficiency [61]. Also, we can count the number of mouse clicks to estimate effort [12] .

R. Sowell [90] listed factors that influence accuracy, consistency, and efficiency while

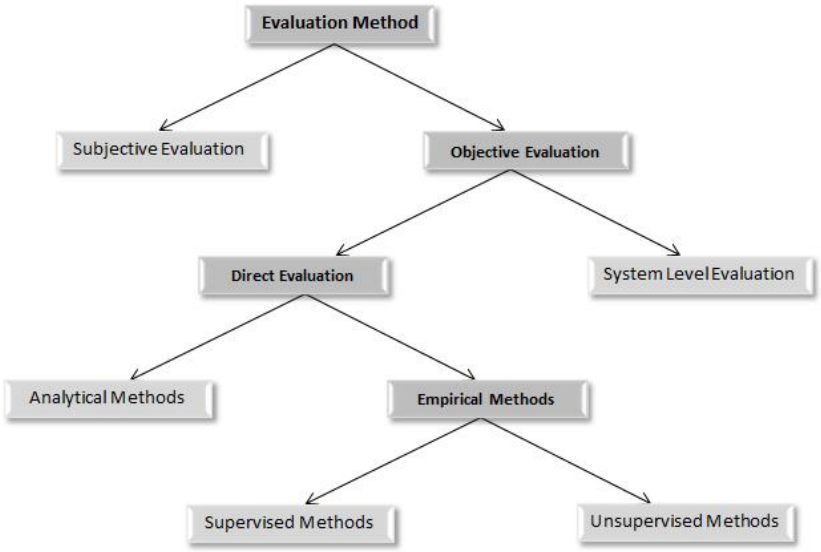


Figure 1.5: Evaluation hierarchy [106]

manually drawing contours during a segmentation. Factors that influence accuracy and consistency are grouped into three categories: differences in judgment, differences in performing the segmentation, and differences in the computational component of the tool. In Table I, we summarize all of these factors.

While there are various methods for image/volume segmentation, the development of systematic evaluation frameworks for segmentation methods is slow. There are some issues that disrupt effective evaluation of a 3D volume segmentation including [95]:

- The data sets are too small
- Different data sets are used for different estimations of performance
- The data sets may not reflect the problem of interest (especially in clinical problems)
- Appropriate ground truths are difficult to define.
- Performance metrics are poorly defined.
- There is poor methodology for training and testing the segmentation algorithms.

Table 1.1: Factors that influence accuracy, consistency, and efficiency [90].

Measure		**Factors**
Accuracy /Consistency	Differences in judgment	Boundary strength: A weaker boundary is more difficult to delineate and cause users to be more likely to disagree.
		Following different guidelines: Users with different knowledge or background follow different guidelines in segmentation.
	Differences in performing segmentation	Shape complexity: It is more difficult to segment a more complex structure.
		Simplification: Users may disagree on how much to simplify a contour.
		User fatigue: A tired user is more likely to make mistakes
		Mental integration: Differences in how users perceive the impact of their contours on the final shape will result in differences in the contours that they draw.
	Differences in the computational component of the tool	Surface reconstruction algorithm: How the contours get interpolated will vary from one algorithm to the next.
		Algorithm parameters: Changing the input parameters of an algorithm will alter the results of segmentation
		Contour filtering: filtering noise from the raw user input will have an effect on the input, and therefore the resulting segmentation as well.
Efficiency/ Elapsed time and user effort		Number of planes: There is an overhead associated with each additional plane that a user must contour.
		Contour length: Longer contours require more mouse clicks and therefore take more time to draw
		Boundary strength: Weaker boundaries require more time for the user to determine where the contour should be drawn.
		Shape complexity: More complex contours require more mouse clicks to capture the details of the boundary and therefore take more time to segment.
		Demand posed on mouse operation: When it is necessary to draw more carefully, the effort to control the mouse increases. It potentially leads to slower operation and higher chance of fatigue.
		Predictability of the method's behavior in response to user input: Users should be able to predict the impact of their actions in results for an efficient interaction.
		Type of knowledge needed to input data: Some tasks, like typing the value of parameters, require knowledge about the working of the algorithm. In this case slower operation is expected.

- Large amounts of time and effort are involved in collecting and hand segmenting data.
- The segmentation algorithms are not compared against other algorithms.

1.2.3 Human Factors in 3D Image/Volume Segmentation

As described in previous sections, substantial work has focused on creating and evaluating various 3D volume segmentation methods but only a few have taken into account humans and their role. [65] presents an early review of human-computer interaction in image segmentation while mentioning three common criteria for segmentation evaluation: Accuracy, repeatability and efficiency (described in Section 1.3.2).

In [54], the author aims to understand how physicians interact with the information in a medical image during the interpretation process. However, medical imaging perception and interpretation is only a part of the 3D volumetric data segmentation process.

In order to assess the quality of expertise differences in the comprehension of medical visualizations [35] reviews quantitative studies that characterize where experts look based on where in the image their gaze rests (fixation) and how long it is there (dwell time). Fixation and dwell time are lower level tasks that relate to higher order tasks of searching, abnormality detection and accuracy of detection. Segmentation, while requiring the ability to detect a feature of interest, also requires the expert to build a 3D structure from those features. Research by [58] lays the groundwork for assessing difference between expert's comprehension of visualization. However, there is no similar work in 3D segmentation area.

The authors of [75] investigate effects of user interaction in semi-automatic segmentation methods for segmenting the organs at risk in radiotherapy planning. Their findings suggest that in the future HCI design of semi-automated segmentation approaches, there should be more flexibility in the interface design, and user interactions should be less cognitively challenging. Unfortunately, they only had two participants in their study which limits inter-observer variations. Their work reflects the mode of data capture, the logic of underlying algorithms, and input devices with the aim of providing flexibility in user interfaces or decreasing the cognitive load of interactions within the overall process. The focus of HCI optimization ultimately arises from the concept of *fewer clicks*, translating to ease of use rather than from the notion of domain knowledge as supportive of ease of

use.

1.3 Research Objectives and Contributions

As 3D volume segmentation is a complex process, it requires expert segmenters to utilize combinations of tool use, workflows, and domain-specific knowledge to complete the process. This process is also highly dependent on human experts as their role in successfully completing an accurate 3D volume segmentation is undeniable.

Our research goal is to investigate 3D volume segmentation as a human-computer interaction paradigm and to understand human factors that are involved in the current segmentation process. Our research addresses the need for methods to understand what experts are doing during 3D volume segmentation, to capture lower-level cognitive and perceptual tasks, higher-level constraints, behaviors and mental models in order to improve the accuracy, repeatability and quality of segmentation. The result of our research is a 2D cross-section training strategy to help both experts and novices learn how to perform an accurate segmentation in a more efficient manner.

We believe by capturing lower and higher level segmentation processes (e.g. visual cues, mark making and placement, predicting correct 2D slides of 3D structures, local criteria for accuracy or quality etc.), and segmenters' mental model of the segmentation processes, we can establish a set of measurable higher-level behaviors, eventually leading to development of design heuristics for segmentation processes, training paradigms, tools, and algorithms to produce those behaviors [102].

Our research contributions include:

1. Design of a hybrid protocol (see Chapter 2: *Formative Study Design*) for formative studies. This protocol design helps us extract tacit expert knowledge within the application domain of 3D volume segmentation [102].
2. Devising a novel coding scheme to capture both low-level (micro) and high-level (macro) segmentation actions and tasks perform by segmenters, and to identify different segmentation approaches utilized by expert versus novice segmenters in the field [81].
3. Developing and validating a 3D spatial ability test instrument to measure 2D cross-section understating skills and evaluate the effectiveness of training strategies for

3D segmentation [79, 80].

4. Designing and implementing a new 2D cross-section training strategy to help novices and people with low spatial skills understand and identify 2D cross-sections of 3D structures in the context of 3D volume segmentation.

1.4 Research Methodology Overview

Figure 1.6 provides an outline of this thesis.

Chapter 2 covers our formative study design that we used to build up our understanding of the segmentation process in the context of expert/novice users performing real segmentation tasks.

In Chapter 3, we analyze data that we captured using the formative study design. We particularly focus on analyzing low-level actions and high-level tasks performed by experts during segmentation.

The research studies presented in Chapter 2 and 3 were completed in collaboration with Dr. Ruth West (professor of University of North Texas), and her undergraduate students team (UNT Team).

Chapter 4 presents our work on designing and developing a new test instrument to measure humans' performance on 2D cross-section understanding tasks. We also introduce our novel range of difficulty for inferring 2D cross-sections and use that in designing and validating question items of the test instrument.

In Chapter 5, we complete developing our 2D cross-section training strategy through the design and implementation of a domain-agnostic and interactive training tool for inferring 2D cross-sections of 3D structures. We used the test instrument introduced in Chapter 4 as a pre/post-test to evaluate the effectiveness of the training strategy by measuring participants' spatial abilities before and after the training.

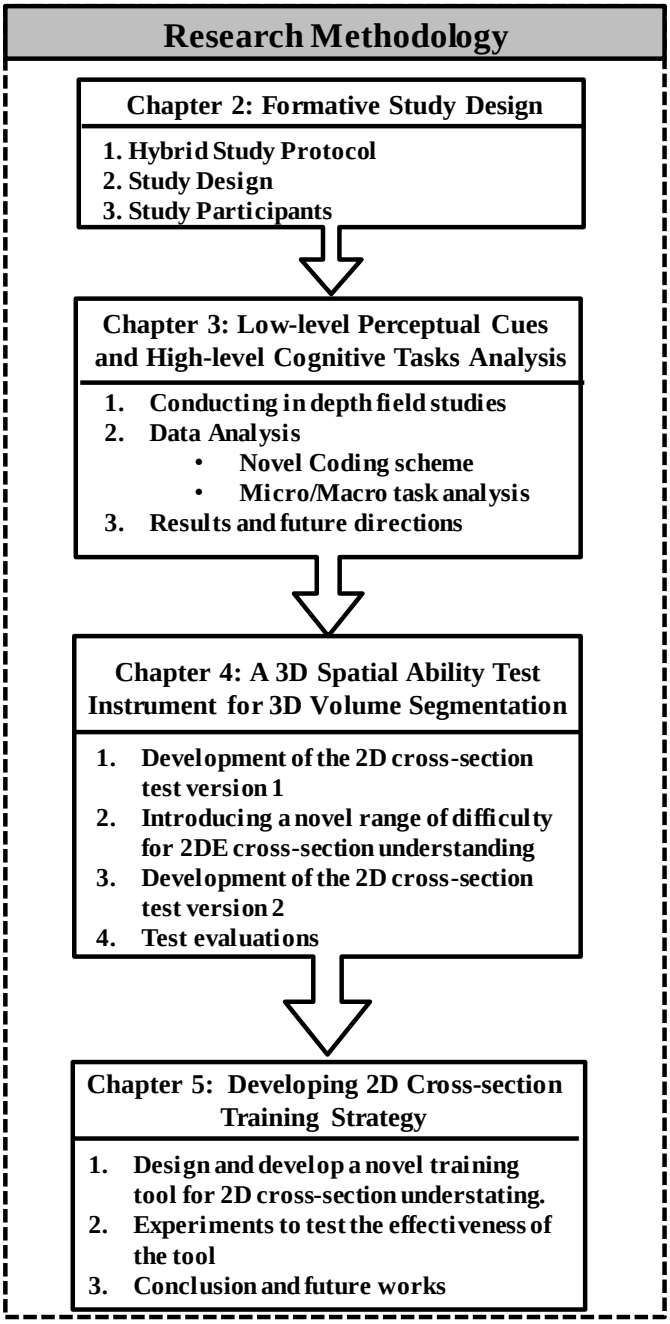


Figure 1.6: Outline of the thesis.

Chapter 2: Formative Study Design

Our research goal is to understand human factors that are involved in a 3D volume segmentation process. Also, we aim to capture enough domain knowledge in order to effectively help segmenters, but not to become experts in segmentation for any particular domain [77, 99]. Understanding 3D volume segmentation process, from human-factors perceptive, is challenging and requires enough attention in both study design and implementation.

Through informal interaction with segmentation experts and novices, we developed an initial understanding of the segmentation process. Based on that, we define 3D volume segmentation to be an iterative process that incorporates several steps including inspection, navigation, marking, editing, and visualization (Figure 2.1). Each of these steps requires various sets of sub-tasks and has its own constraints and challenges.

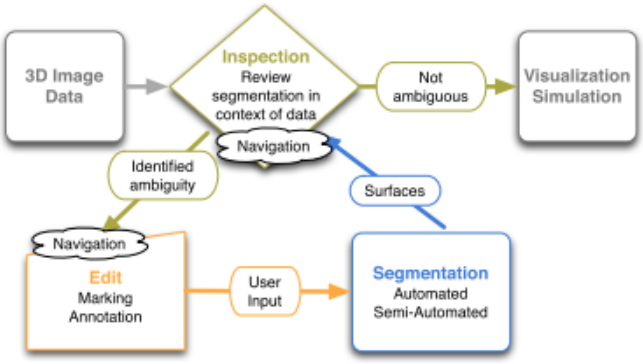


Figure 2.1: Integrated pipeline of segmentation, inspect, and edit.

We aim to enhance our understanding of each step of the segmentation process. But the question is how? To do that, we propose to conduct formative studies to understand the segmentation process in the context of experts performing real segmentation tasks. Initially, we chose field studies and in-depth, per-expert analysis over specific controlled testing because the tool, data sets and expert approaches are very idiosyncratic, and

running a formalized controlled study too early risks missing non-obvious elements or biasing our results.

The first step before conducting field studies is to design a solid study protocol. To address different characteristics of segmentation, our study protocol benefits from existing qualitative and quantitative human-computer interaction methods so that it can be utilized in the field for formative studies and later combined with additional controlled ones. More specifically, our study protocol is modeled after *Design Studies* [85] in visualization, which has close ties to *Contextual Inquiry* [44], *Action Research* [38], and the emerging field of *Visual Analytics* [4].

Our field studies take place at segmenters' site (e.g., their lab), while they are working in their own environment, using their own tools and computers. To capture segmenters' tasks and behaviors as much as possible, we carefully observe them, and ask our semi-structured interview questions. We record data using different methods including Eye-tracking (ET), over the shoulder video (OTS) and audio recording.

Our protocol builds upon our knowledge of the problem domain and its design aims to capture expert knowledge and mental model through observations in the field while segmenters are using their own tools and datasets. We need to have a solid study design to help us run our field studies. On the other hand, to complete designing our study protocol and make sure it works correctly, we need to implement it in real world through our field studies. To solve this problem, we introduce our hybrid protocol which consists of two main designs:

1. **Initial study design:** This is our preliminary design which provides guidelines and instructions to run a field study in three main steps of pre-observation, main-observation, and post-observation. The goal is to gather in-depth data through eye-tracking, and video/audio recording. After conducting a few field studies and analyzing the results, we understood this design has some limitations. Therefore, we tried to overcome these limitations in our revised study design. In Section 2.1.1 and 2.1.2, we talk more about this design and its limitations.

2. **Revised study design:** After conducting a few field studies and analyzing the results, we identified limitations of our initial study design. To address these limitations, we revised our protocol. Our revised study design combines ethnographic *Semi-Structured Interviews (SSI)* with *Retrospective Think-aloud (RTA)* [28, 46],

Cognitive Task Analysis (CTA) [17], *Critical Decision Method (CDM)*[53], and *Eye Movement Tracking (ET)*. For full description of revised design see Section 2.1.3. In the following, we briefly describe each of these methods:

- **Semi-Structured Interview (SSI):** Semi-structured interviews utilize a set of themes to guide the conversation across relevant topics without interrupting the flow [10]. We use SSIs in the pre and post interview stages in order to ensure that we cover all topics related segmentation.
- **Retrospective Think-aloud (RTA):** RTA is a form of think-aloud protocol performed after the task session and has users provide a task analysis (TA). Usually RTA protocol is stimulated by using a visual reminder such as a video replay. We utilize eye-gaze overlay to stimulate the RTA [26, 46].
- **Cognitive Task Analysis (CTA):** The CTA protocol is based on "cognitive task analysis" using the "concepts, processes, and principles" model [17, 18, 19]. It goes beyond standard task analysis methods to capture knowledge, thought processes and objectives that underlie tasks performed.
- **Critical Decision Method (CDM):** CDM is an ethnographic method that reduces reliance on individual memory by guiding the participant through a series of multiple passes of event retrospection [53]. CDM seeks to identify cues and other information that enables experts to identify critical events and informs their decisions at these points. We use CDM to elicit higher-level decision making not apparent in the CTA.
- **Eye-tracking (ET):** Eye-tracking involves measuring either where the eye is focused or the eye-movement [35]. It helps us understand what experts are looking at during the process of 3D volume segmentation.

Our field studies focus on formally documenting the 3D volume segmentation process with a broad range of domain experts on actual segmentation tasks. So, our study protocol is designed in away to help us achieve the goal of formalizing low-level and higher-level actions, tasks, and constraints of the 3D volume segmentation; also to elicit segmenters' behaviors and mental model, while providing a mechanism to compare novices with experts.

In this chapter we present our formative study protocol, which we call "Hybrid Protocol", in more details.

2.1 Hybrid Protocol

To replicate segmenters' normal process as much as possible, we need to observe them in their work settings, using their own tools, data, and computers. Therefore, we chose field study as our formative research method to investigate 3D volume segmentation process as a contemporary set of events over which we have little or no control [104]. Our field protocol is designed for observing participants (experts and novices) in their work settings, and it seeks to elicit tacit expert knowledge through the observations in the field. The aim is to not only understand segmentation process, but also how novices differ from expert segmenters.

Through informal interaction with segmentation experts and novices we developed an initial understanding of the kinds of limitations and problems encountered in the field. This helped us coming up with our initial study design. This design is basically includes field observations and interviews, and is conducted in three stages of pre-segmentation, segmentation, and post-segmentation. The eye-tracking (ET) data along with over the shoulder video (OTS) recording are the main sources of data.

We used initial protocol to run two field studies in two different sites. After analyzing the results, we identified its limitations and to address them we developed our revised design.

The revised design is a more complicated version of the initial design which combines ethnographic Semi-Structured Interviews (SSI) with Retrospective Think-aloud (RTA), Cognitive Task Analysis (CTA), Critical Decision Method (CDM), and Eye Movement Tracking (ET). The Naturalistic Observation (NO) eye-tracking data and scene camera video along with NO over the shoulder video serve as the foundation of the protocol.

We combine RTA and CTA, and use them to reveal conceptual and procedural knowledge, cognitive processes; also to evaluate segmentation outcomes. Low-level actions can be captured through video/audio and ET data. We use replay of the eye movement video and CDM to elicit from participants links between lower-level (visuals and marks) and higher level (perceptual and cognitive) tasks and behaviors. Using observation, eye-tracking combined with RTA, a prompted multi-pass recall structure similar in format to CDM, and semi-structured interviews help us create linkages between thought processes and observable behaviors.

Analysis and results of studies with this protocol, shows its effectiveness in both

capturing low-level actions/ higher-level tasks, and eliciting experts behavior during a segmentation process (See Chapter 3 for full analysis and results).

Figure 2.2 shows our hybrid protocol in a glance.

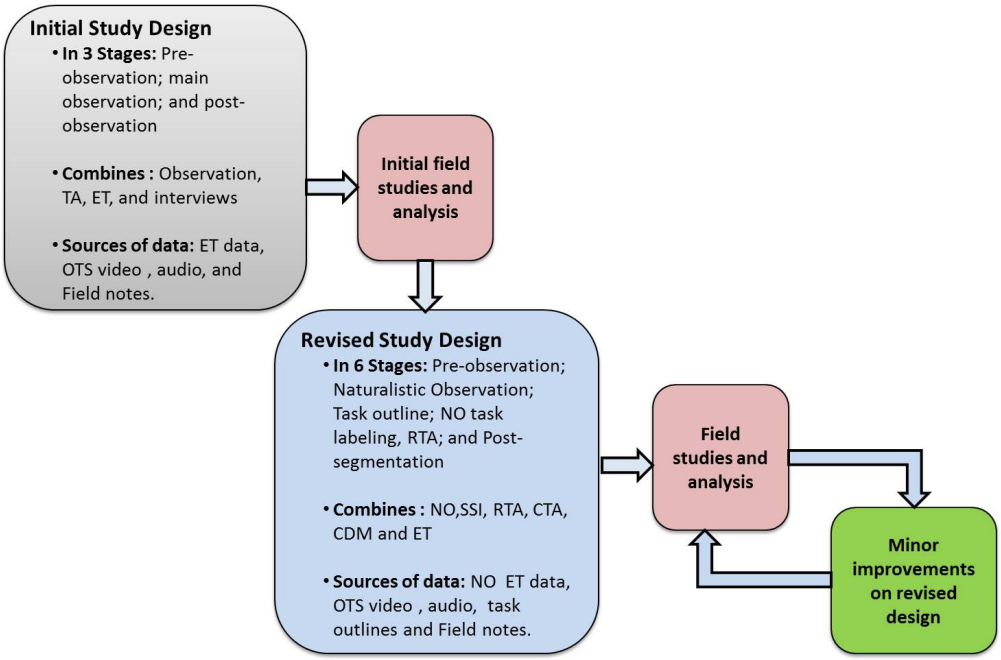


Figure 2.2: Hybrid protocol: Initial and revised design

2.1.1 Initial Study Design

In our initial study design, to capture segmentation process as much as possible, we have a set of questions to ask from participants. These questions are reflected in three stages of a study: a set of pre- and post-observation questions (see Appendix item "Segmentation Study Materials"), and the main observation session where participants used their segmentation tool. We use different sources for gathering the data at each stage as shown in Table 2.1.

Using initial protocol, the video is captured by placing a video camera over the left shoulder of the participant in order to capture the screen and mouse. The eye-tracking utilized a 500 MHz SMI remote tracker. We record audio both through the video camera,

separate microphone, and the eye-tracker device (we merged the over the shoulder video with the eye-tracking data to achieve higher quality video), to make sure the video and eye-tracker data are synchronized. Study sessions last on average an hour, with a maximum of 2 hours. The on-site study team consist of 3 researchers:

- A primary interviewer who asks the questions.
- A secondary interviewer who is responsible for quality control and follow up questioning.
- A note taker who also captures video, sound and eye-tracking.

In the following, we describe each of the stages:

1. **Pre-observation Stage:** At this stage, we ask warm-up questions to relax the participants and get an approximate idea of how and why the participants are segmenting the data. Specifically, we ask questions to determine the domain and sub-domain of the segmentation task, how participants acquire the volume data, any pre-processing techniques used, the goal or output of the segmentation, and the participant's experience level. Because this stage does not involve working with the tool or performing any segmentation, we only record audio. Hand-written notes are another source of data for triangulation, helping us to organize the information before annotating the audio and video.
2. **Main Observation Stage:** During the observation step, participants are asked to think-aloud while doing their segmentation tasks as close to their normal process as possible. For longer, or repetitive, segmentation tasks participants should demonstrate examples of how they performed different stages of the segmentation task.

Table 2.1: Initial protocol data sources

Stage	Eye Tracking	Video (Camera)	Audio (Mic.)	Audio (Camera)	Audio (Eye Tracker)	Hand Written Notes
Pre-observation			Yes			Yes
Observation	Yes	Yes	Yes	Yes	Yes	Yes
Post-observation		Yes	Yes	Yes		Yes

3. **Post-observation Stage:** At this stage, questions cover high-level strategies and mental models, repeatability measures, post processing and what necessitates redoing a segmentation. Participants are free to use their segmentation tools and other downstream applications to answer these questions.

2.1.2 Why Do We Need a Revised Study Protocol?

Using our initial study design, we conducted two field studies. During data analysis for these studies, we identified the below problems and limitations that the initial study could not address:

- **Think-aloud and eye-gaze disruption:** As cognition and perception are linked [35], in studies where eye-tracking is required, the use of a think-aloud protocol while capturing eye movement, can negatively affect the participant's eye motion and locus of attention during task performance. We realized that our participants also did not look at the place we expected them to look (e.g. the region of the structure) while thinking aloud. Therefore, we could not use all the observation period data for our analysis. For example, for one participant we ended up only using 20 minutes of data, while the whole observation was around 60 minutes.

  It is necessary to have an uninterrupted segmentation time, without the participant talking or interacting with the experimenters in any way. In our revised design, we overcome this problem by separating think-aloud from observation using NO and RTA (See Section 2.1.3 for details).

- **Fail to capture links between lower-level tasks and higher-level behaviors:** In our initial study design, no mechanism was provided to help us elicit the link between lower-level (e.g., visuals and marks) and higher level tasks and behaviors. (perceptual and cognitive). To address this issue, in our revised design, we use replay of the the eye movement video and CDM.

- **Too many questions were asked in interviews:** Initial field studies analysis brought our attention to the fact that we were asking too many questions in our interviews. Some of these questions did not add value to our analysis and just made our participants tired and confused. Additionally, despite of asking so many

questions, we failed to elicit the hows and the whys of participants decision making. Obviously, we need to change the questions in our revised protocol.

- **Problems with eye-tracking device:** In initial studies we used under-the-monitor eye-tacker. Unfortunately, the eye-tracker did not work as we expected. One of our problems was the calibration process. The eye-tracker forced users to not move their head and even small head movements could lead to re-calibration. In our revised design, we suggest using glasses which gave participants enough freedom to move their heads without a need to redo calibration. It also provides higher quality output video.

2.1.3 Revised Study Protocol

Using the revised protocol, a field study session duration is 1 to 2 hours. Similar to the initial study design, the observation set up includes over-the-shoulder video, audio recording, hand-written notes, and eye-tracking using an SMI remote 500 Hz glasses based tracker. As mentioned previously, we replace under-the-monitor trackers with glasses to give our participants freedom to move their heads without a need for re-calibration. Figure 2.3 shows observation environment.

Three roles are defined to facilitate each field study:

- Primary Interviewer who conducts the interview and observation.
- Secondary Interviewer who tracks topic coverage and asks follow-on questions.
- Note taker who takes note, and captures video, audio and eye tracking.

The protocol consists of 6 main steps as depicted in Figure 2.4:

- **Step 1 (Pre-Observation):** After giving the consent form to participants and describing research goals to them, step 1 starts with semi-structured interviews (SSI). We ask participants to describe their segmentation data and tools, and provide us with an overview of the whole process.
- **Step 2 (Naturalistic Observation):** In Naturalistic Observation (NO), participants perform as much of their complete segmentation process as naturally as possible while skipping repetitive actions after a few examples, and without talking, We use video recording and eye-tracking at this phase. Calibration of the eye

Figure 2.3: Observation environment.

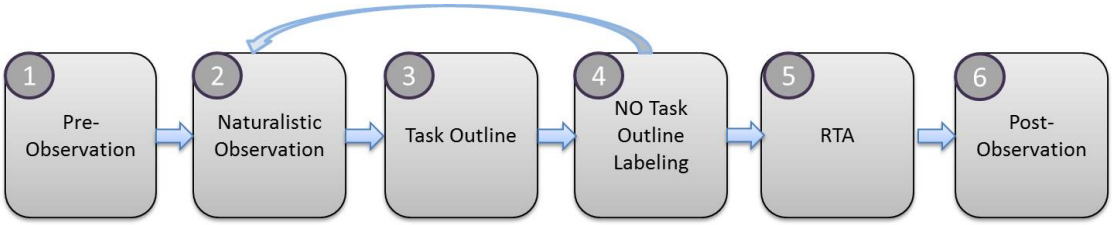


Figure 2.4: Revised study protocol.

tracker occurs prior to the NO. We monitor the eye tracking during data capture and re-calibrate as necessary upon detection of drift using times when the gaze should match the cursor location as a guide.

- **Step 3 (Task Outline):** In this step, we ask participants to hand-write a task outline (TO) on paper that describes the steps of their segmentation process.
- **Step 4 (NO Task Outline Labeling):** This step combines CTA and CDM. Participants label their NO eye-tracker scene video according to the task outline completed in the prior phase. If necessary, they modify their hand-written CTA. We capture more NO video to track the modifications, if needed.

- **Step 5 (RTA):** In this step, we ask participants to describe what they were thinking as they viewed (playback) or scrubbed through the NO video with eye-gaze overlay.
- **Step 6 (Post-Observation and Conclusion):** Using semi-structured interviews we go in depth on the participant's decision making process. Finally we conclude with a participant debriefing.

We utilize over the Shoulder Video (OTS) with audio throughout the protocol. We capture eye-tracking data in step 2, and if needed in steps 4 and 5. During steps 4 and 5, we record the playback of the NO video in order to synchronize the OTS video and audio with where participants were looking in the NO video. Table 2.2 shows different sources of data for these steps.

Table 2.2: Revised study protocol data sources. * indicates eye-tracking is used if needed.

Stage	**Eye Tracking**	**Video (Camera)**	**Audio (Mic.)**	**Audio (Camera)**	**Audio (Eye Tracker)**	**Hand Written Notes**
Pre-observation		Yes	Yes	Yes		Yes
Naturalistic Observation (NO)	Yes	Yes	Yes	Yes	Yes	Yes
Task Outline		Yes		Yes		Yes
NO Task Lableing	*Yes	Yes	Yes	Yes	*Yes	Yes
RTA	*Yes	Yes	Yes	Yes	*Yes	Yes
Post-observation		Yes	Yes	Yes		Yes

As explained previously, the Naturalistic Observation (NO) eye tracking data and scene camera video along with NO over the shoulder video serve as the foundation of the protocol. The blend of direct observation, eye motion tracking combined with RTA, a prompted multi-pass recall structure similar in format to CDM, and semi-structured interviews enables us to conduct better field studies to capture lower-level/higher-level segmentation tasks and actions, and create linkages between thought processes and observable behaviors.

However, CDM requires a list of questions to elicit the hows and the whys of participants decision making. In order to address the needs of CDM, and to ensure that

we were as uniform as possible in our data capture across all participants, we developed a set of codes and corresponding prompts encompassing segmentation processes end-to-end. These codes were developed from our experience in the field and refined after each observation. We refer to these codes and prompts as our *Master Annotation Codes (MAC)*.

MAC is subdivided into 16 codes that span from 3D data capture, through segmentation, to downstream uses of segmentation products. These codes include: *Image Data Acquisition*, *Data Pre-processing*, *Process Strategies*, *Data Exploration*, *Marking Stage*, *Automated Stage*, *Review/Refinement Stage*, *Data Post-Processing*, *Downstream Use*, *Overall Process*, *Data Relation to Use Case*, *Quality Evaluation in Use Context*, *Navigation Strategies*, *Surface Understanding*, *Tools and Quality Control*. The MAC is used to create a series of prompts for CDM and pre-observation steps. It also helps us organize our interviews, comping up with more meaningful and concise questions.

After running and analyzing a few more field studies with revised protocol, we verified that the protocol, with only a few minor improvement, was solid enough to be used in all of our remaining field studies (See Chapter 3 for the field study analysis and results).

2.2 Participants and Dataset

We selected 10 participants from 5 different sites recruited through personal contacts. The number of participants per site ranged from 2 to 3. Tools ranged from custom-built to commercially available software. These participants covered a range of tasks, data sets, tools, and expertise. Because segmentation is a complex task that involves different visual search skills and domain knowledge, we believe that using a small number of experts that are well-trained in the respective areas of the tasks is warranted [9, 20]. We balanced the small number of experts in each domain by using several diverse domains and tools. We expect our participants to be representative of segmenters who are working on 3D image segmentation projects. Table 2.3 shows participants of our field studies. Also, Figure 2.5 shows some of the data sets that participants worked with.

Table 2.3: Participants of the study

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Expertise	Radiology	Biomedical	Soil Analysis	Cell Biology	Cell Biology	Cell Biology	Cell Biology	Cell Biology	Cell Biology	Cell Biology
Years Segmenting	>5 yrs.	>3 yrs.	>1 yr.	<1 yr.	<1 yr.	<1 yr.	>5 yrs.	>1 yr.	>3 yrs.	>40 yrs.
Prior Experience with Data Set	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Site	Site 1	Site 1	Site 2	Site 2	Site 3	Site 3	Site3	Site 4	Site 4	Site 4
Tool	Tool 1	Tool 1	Tool 2	Tool 3	Tool 3	Tool 3	Tool 3	Tool 4	Tool 4	Tool 4

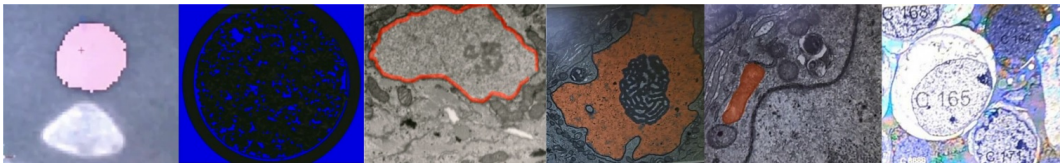


Figure 2.5: Data set examples. Left two: CO^2/soil analysis, right four: cell analysis.

2.3 Conclusion

In this chapter, we presented a hybrid protocol design for capturing low-level actions, higher-level tasks, and 3D volume segmentation task work-flow of segmenters, in the field. The protocol consists of initial and revised design. Revised design was introduced to overcome the limitations of initial design (e.g, eye-gaze and think-aloud disruption). It combines ethnographic Semi-Structured Interviews (SSI) with Retrospective Think-aloud (RTA), Cognitive Task Analysis (CTA), Critical Decision Method (CDM), and Eye Movement Tracking (ET). The Naturalistic Observation (NO) eye-tracking data and scene camera video along with NO over the shoulder video serve as the foundation of the design. We used the revised study design in our formative field studies (Chapter 3).

Chapter 3: Analyzing 3D Volume Segmentation by Low-level Perceptual Cues and High-level Cognitive Tasks

We conducted formative field studies to further understand the segmentation process in the context of experts performing real segmentation tasks, and to capture both low-level (micro) and higher-level (macro) tasks during 3D volume segmentation. It may seem obvious that people look at what they are segmenting, but we do not know the specific image features experts use, how experts determine their segmentation strategies and approach, and how that might vary during segmentation. Additionally, we do not know the differences between experts and novices in performing 3D image segmentation.

In this chapter we investigate how experts do 3D volume segmentation by quantifying their low-level actions, tasks and behaviors during the segmentation process. Figure 3.1 shows an overview of different parts in analyzing 3D volume segmentation by low-level perceptual cure and high-level cognitive tasks.

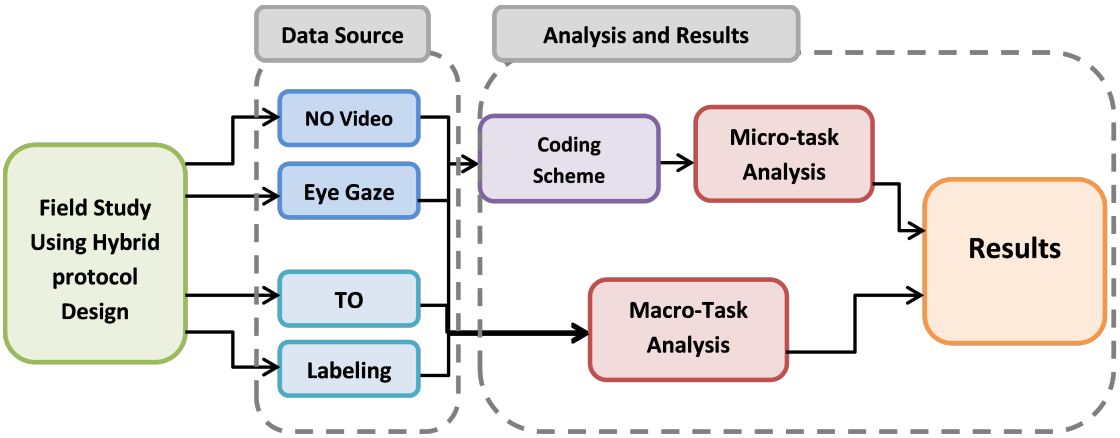


Figure 3.1: Analyzing 3D volume segmentaion process.

In this chapter we cover the followings:

- **Overview of field studies using our hybrid study protocol:** Using our field studies, we captured as much data as possible to be used for analyzing the

segmentation process. Our sources of data include observation video/audio, eye-tracking data, task-outlines created by our participants, and field notes.

- **Details of our data analysis methodology:** In this chapter, we explain our analysis methodology for cleaning, organizing and analyzing the field study data.
- **Details of our results:** After analyzing our data, we present our results and insights using different figures and tables.

Research Objective: Identify and classify experts' eye-gaze location, low-level actions and tasks during 3D image segmentation (we split action and tasks into two categories of micro and macro tasks). In Section 3.2 we explain these categories in more details.

Our main research questions are:

RQ1: Can we identify where segmenters look and what their low-level (micro) and higher-level (macro) segmentation actions and tasks are?

RQ2: How do expert and novice segmenters' gaze patterns differ for different/similar segmentation micro/macro tasks?

Contributions:

1) Devising a novel coding scheme that encodes segmenters' observable, low-level actions and what they are looking at when performing those actions; 2) creating a domain-agnostic classification to capture segmentation-level tasks; and 3) defining different segmentation approaches and strategies used by participants.

3.1 Field Study: Capturing Micro and Macro Segmentation Tasks

There exists little knowledge regarding experts' low-level perceptual (micro) and higher-level cognitive (macro) tasks in the segmentation process. To address this, we choose to conduct field studies and in-depth, per-expert analysis over specific controlled testing because the tool, data sets and expert approaches are very idiosyncratic, and running a formalized controlled study too early risks missing non-obvious elements or biasing our results. We used our hybrid protocol (See Chapter 2: *Formative Study Design* for details) as a guideline to run the field studies.

We conducted two field studies using our initial study design. After preliminary analysis, we managed to identify the design limitations (explained in Chapter 2, Section 2.1.2). we addressed those limitations through combining ethnographic Semi-Structured

Interviews (SSI) with Retrospective Think-aloud (RTA), Cognitive Task Analysis (CTA), Critical Decision Method (CDM), Eye Movement Tracking (ET), and Naturalistic Observation (NO). We then conducted three more field studies using the revised design. Further analyses, which are explained in this chapter (see Section 3.3), show that the revised deign is promising. Therefore, we use this design for any additional field study.

In the following subsections we talk more about micro and macro tasks, sources of data, and our participants.

3.1.1 Micro/Macro Tasks Definition, and Sources of Data

As explained in Chapter 2, our study designs incorporate different sources of data as:

- In initial study design, we used data of pre/post and observation stages. Data is captured using eye-tracking, video/audio recording and field notes.
- In revised study design, we utilized data from pre-observation, naturalistic observation (NO), participants' task-outline and labeling, RTA, and post-observation. Again, we used eye-tracking, video/audio recoding (including over the shoulder video), and field notes to capture data.

Using the above sources of data, we captured two distinct categories of segmentation tasks, which we call micro- and macro- tasks. Broadly speaking, micro-tasks are low-level actions that can be observed in the eye-tracking video, while macro-tasks come from the participant's cognitive task outlines. These are summarized in Table 5.2.

3.1.2 Participants

As we described in Section 2.2, in the field studies we observed 10 participants. Table 2.3 summarizes the specification of these 10 participants. We chose these participants for further analysis because they covered different range of tasks and data sets. Also, the data set they worked with was higher quality, and we had more reliable eye-tracking data for these participants.

P1-P4 are participants of the field studies with the initial study design, while P5-P10 are participants of the field studies with revised design.

Table 3.1: Micro/Macro Task Definition

Category	Micro-Task	Macro-task
Definition	Low-level actions and tasks accomplished by segmenters which are observable in an observation/NO video. We consider these actions something we can see in the video, not something we assume participants are doing.	Higher-level segmentation tasks which naturally consist of smaller sub-tasks and low-level actions. We identify these tasks through the CTA protocol and participants' task outlines
Goal	To identify where segmenters' are looking during segmentation, and to quantify lower-level actions and tasks of participants in a segmentation process.	To identify what participants are doing during different stages of the segmentation process, and what are the domain-agnostic phases and patterns of higher-level actions during segmentation.
Data Source	Observation (initial study) and NO (revised study) video (including eye-tracking video and OTS)	Combination of participants handwritten task-outline, task labeling, RTA and CDM
Example	Looking at the tool or image, paging, zooming and drawing a contour	Data initialization, inspection, working with 2D and 3D view, and reviewing

3.1.3 Eye-tracking Data Cleaning and Gaze Quality

Eye-tracking data is an important source of data in our research. So, it is necessary to capture eye-gaze data as accurately as possible. To make eye-gaze data usable, we create gaze-overlays which combines eye-gaze raw files with eye-tracker video.

In our initial field studies we used under the monitor eye-trackers. Unfortunately, in this case, calibration was challenging as participants sometimes moved their heads and we ended up to redo the calibration several times. We constantly monitored the tracking during the study capture and re-did the calibration as necessary when it drifted, using times when gaze should match cursor location or button (e.g. adjusting a slider). We re-checked this calibration before analysis, excluding two participant segments in the initial studies.

In initial studies, to create gaze/heatmap overlays, we used a python script along with our OpenCV program. This also raised problems as sometimes the raw data was not synced with the eye-tracking video, or had low quality which forced us to pair over the shoulder videos with eye data.

To solve these problems, in our revised studies, we replaced the under-the-monitor eye-tracker with glasses to give our participants more flexibility to move their heads. We used the software supplied with the eye tracker to create the overlays. For calibration, we asked participants to look at four different places in their monitor (marked with four

stickers). We re-did calibration only a few times.

While segmenting, the participant's gaze primarily consisted of smooth pursuit along identifiable features in the image data or transitions from a tool menu on the side to the data; therefore we did not have to make judgments about rapid saccades within a small region or try to determine which of many small features the gaze overlapped.

Eye-tracking videos of glasses were shaky due to head movements. We used the stabilization approach in OpenCV to stabilize all videos before using them in our analysis.

3.2 Data Processing and Analysis

The Naturalistic Observation (NO) eye-tracking data and scene camera video along with NO over the shoulder video serve as the foundation to capture micro-tasks. Macro-task capturing is accomplished using task-outline, task labeling, RTA, and CDM. We define our analysis process as following:

1. Combine NO videos with Heatmap/gaze overlays to determine participants gaze location. We utilize these data to capture micro-tasks.

2. Create a novel coding scheme to categorize micro-tasks. We assign a code to each identified micro-task. The coding scheme is in a hierarchical format.

3. Modify hand-written task-outlines to be simplified or extended. Then, create a mechanism to classify macro-tasks based on the modified task-outlines.

4. Transcribe all the video/audio based on what segmenters were saying during the study.

5. Micro-task analysis based on coding scheme to count the code frequency for each of the tasks. This step validates our coding scheme from step 2.

6. Macro-task analysis using modified task-outlines to identify what participants are doing during different stages of the segmentation process.

7. Tool-feature/data/strategies analysis to classify what participants were saying about the segmentation process.

3.2.1 Micro-Task Coding Scheme

The goal of developing a coding scheme is to support analysis and classification of participants' micro-tasks including perceptual cues, low-level actions (e.g. drawing) and tasks (e.g. marking vs. low-level reviewing). To our knowledge, the research literature does not report coding schemes that are directly applicable to 3D segmentation.

We consider a micro-task to be something we can see in the video data rather than actions we think or assume the participant is doing. We associate a dedicated code to each task. Micro-tasks naturally formed a hierarchy and are grouped into five higher-level action types: 1) Eye-gaze location: participant's eye-gaze location; 2) Global navigation: navigating between slices. 3) Local navigation: zooming and panning; 4) Marking: drawing or editing a contour; and 5) Review: reviewing a segmentation process or mark. Table 3.2 shows overview of our coding scheme. We now briefly describe our coding instructions, and how we used the coding scheme to capture micro-tasks.

Eye-Gaze Location, Data and Tool: The primary objective of these codes is to identify when the eye-gaze is on the tool/UI and when it is on the data. To be considered "Data", the eye-tracker gaze overlay must be completely inside the data/image plane, or at a minimum it is touching the outside edge of the data/image plane. To be considered "Tool", the gaze overlay should completely be inside the tool UI, or at a minimum it is touching the outside edge of the tool UI. Figure 3.2 shows an example for "Data" and "Tool".

Gaze Location Codes, Boundary, Region, Non-ROI features, and Movement: The primary objective of these codes is to identify where in the data the eye-gaze is pointing. These codes only occur within "Data" codes. "Region", "Boundary", and "Non-ROI" signify whether the gaze is within the region of interest, on the boundary of the region of interest, or somewhere else. "Movement" occurs when the eye-gaze shifts from the data to the tool/UI. To be considered "Boundary", the eye-tracker gaze overlay must be on the edge of the ROI or must be touching the edge either inside the ROI or outside. To be considered "Region", the gaze overlay should be inside the region but not touch the edge inside the ROI. Figure 3.3 shows examples for "Boundary" and "Region" codes. To be considered Non-ROI Feature, the gaze overlay is anywhere in the data that

Table 3.2: Micro-Task Coding Scheme

Action Type	Code	Description
Eye gaze Location	Data	User's eye gaze is completely inside the data/image plane itself or touching the edge.
	Tool	User's eye gaze is inside the tool UI, or at least touching the outside edge of the frame
	Boundary	User's eye gaze is in contact with the edge of the region of interest.
	Region	User's eye gaze is fully within the edge of the region of interest
	Non-ROI Features	Eye gaze in anywhere in the data image that is not the region of interest or its boundary.
	Movement	Eye gaze while it is in motion from Tool to Data
Global Navigation	Page	User navigates between slices/planes of the data.
	Free Rotate	User rotates data volume freely in x,y,z axes
Local Navigation	Pan	User drags view of data along x or y axis
	Zoom	User expands view over a portion of the data
Marking Action	Draw	User manually creates a mark by dragging their cursor, usually to encapsulate a figure or draw a line.
	Mark	User creates a special mark (e.g. circles) using the segmentation tool.
	Fill	A region of the data/image plane is filled in
	1st Mark	Denotes the creation of the first mark in a data set and the processes leading up to it.
	Commit	A created mark is saved or approved as done by the user in the program
	Edit	Changes made to an existing mark
	Delete	Mark is removed
Review	Mark Review	Eye gaze lingers on a completed mark
	Assess Data	Eye gaze is on the data. Does not occur during marking action or when looking at a mark
	Copy	Propagation of the settings or marks from one slice to a different slice
	Tool Output	The output that the tool generates.

is not region or boundary (e.g. not touching the ROI and still within the data/image plane. "Movement" code captures when the eye-gaze is transitioning from one window (tool or data) to the other. This code is to differentiate between when the expert is actually looking at the data or tool and when their eyes are on a window because they are looking back and forth between windows/screens.

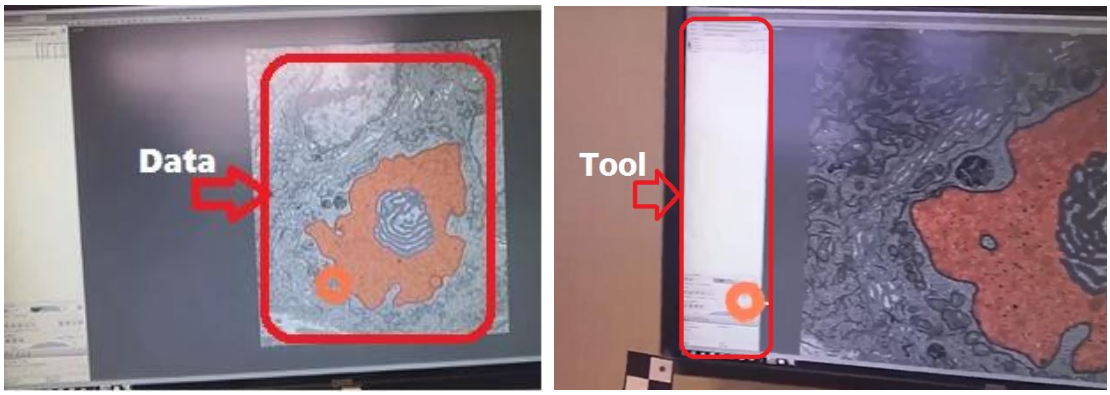


Figure 3.2: Eye-gaze location examples for Data (left) and Tool (right). The small orange circle is the eye-gaze overlay.

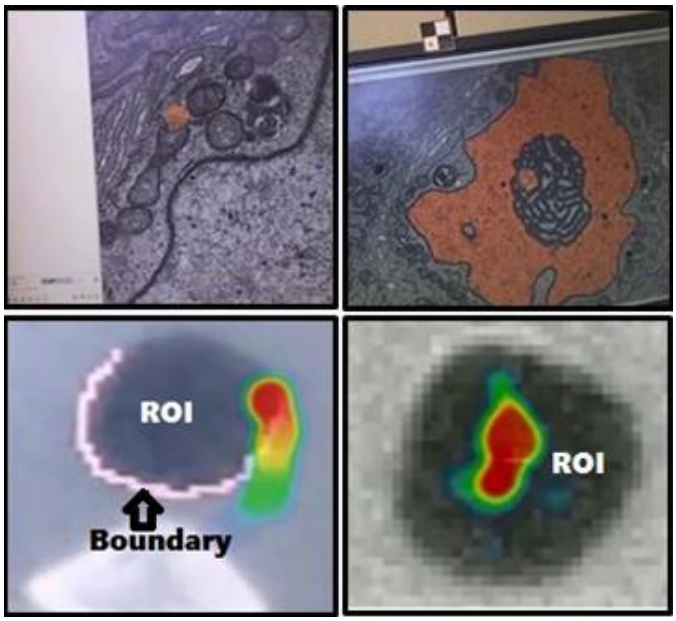


Figure 3.3: Eye-gaze location examples for "Boundary" (top left and bottom left) and "Region" (top right and bottom right). The orange circle (in top images) and heatmaps (in bottom images) are the eye-gaze overlays.

Mark and Navigation Codes: The objective for these codes is to identify marking and navigation micro-tasks during segmentation. These codes may occur across both

"Tool" and "Data" codes because some of these tasks are completed through interacting with the tool/UI. Navigation codes like "Page" are easier to identify in the NO video, so it is more efficient to initially annotate video using these codes and then continue with marking codes. Figure 3.4 and 3.5 show examples for "Draw" and "Fill" codes.

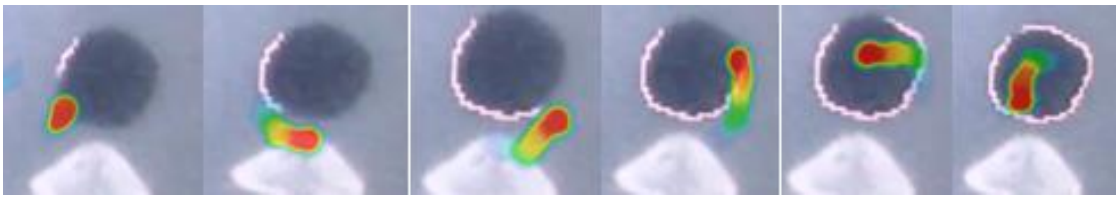


Figure 3.4: "Draw" Code Example. Participant captures the structure by drawing (pink circle). Their heatmap gaze overlay follows the pink circle.

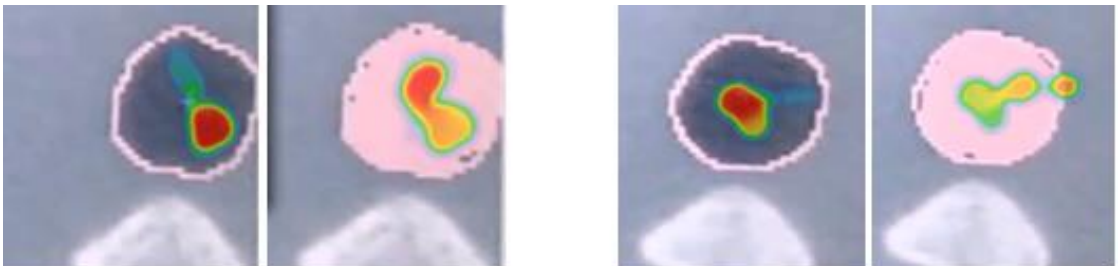


Figure 3.5: "Fill" Code Example. Before and after Filling the structure. The gaze overlay is inside the region of interest.

Review and Assessment Codes: The objective of these codes is to identify micro-tasks attached to some segmentation thought and action processes. These codes belong under the "Data" code. "Assess Data" occurs when the eye-gaze is on the data, but a mark has not been made yet. "Review Mark" occurs when the eye-gaze is on a mark or commit that has been made. "Tool Output" occurs when the segmented slices are put together to create a 3D representation/model of a structure.

3.2.1.1 Micro-Task Coding Procedure and Agreement

We used our coding scheme to annotate observation/NO videos with gaze overlays. To code and annotate our data, we used MAXQDA, an application which supports qualitative data analysis [55]. To code data, we reviewed the videos with gaze overlays. We used the codes in our coding scheme (Table 3.2) along with our coding instructions to complete the annotation process. Coding procedure was accomplished in two steps:

Step 1 (Coding of Initial Studies Data): Unfortunately, in our initial studies, use of a think-aloud protocol while capturing eye-movement negatively affected the participants eye motion and locus of attention during task performance. Therefore, not the whole duration of observation videos were usable for coding. To effectively code our data, we started by making a pass (**Pass 0**) through videos to identify:

1. All those places in videos were participants's eye-gaze was not affected by think-aloud (usable data). We coded those places as *Natural Segmentation*.
2. All those places in videos were participants's eye-gaze was negatively affected by think-aloud (unusable data). We coded those places as *Unnatural Segmentation*.

Initially, three of the researchers independently coded the data using MAXQDA. They annotated observation videos for participants of the initial studies (P1-P4). They started with Pass 0 to identify *Natural Segmentation* parts in videos. They continued the coding in four subsequent passes for only those places in videos which had a *Natural Segmentation* tag:

1. **Pass 1:** *Tool* and *Data* codes.
2. **Pass 2:** *Boundary*, *Region*, *Non-ROI*, and *Movement* codes.
3. **Pass 3:** *Navigation* and *Marking* codes.
4. **Pass 4:** Segmenters Review codes.

Coding agreement was achieved by viewing all three reviewers' codes on a single video. Each researcher started coding by going through a pass. There was an inter-coder agreement after each pass. Inter-coder agreement was calculated using MAXQDA (agreement in percentage). Once all the coders finished a pass and uploaded them to

a shared drive, they used MAXQDA, to check inter-coder agreement. If agreement was <80%, researchers reviewed codes to identify trouble spots and recode if necessary. After iterating on codes and reaching the agreement (for P1 and P2 data), the code set was considered to be robust, and two of the researchers coded the remaining videos, separately (for P3 and P4 data).

Step 2 (Coding of Revised Studies Data): During initial study coding, researchers became familiar with the coding procedure and achieved enough agreement on coding. The procedure is almost the same for the revised study data. The only exception is, since we were annotating NO videos, we did not need to identify any Natural/Unnatural Segmentation parts in video (no need to go through Pass 0).

We decided to check inter-coder agreement for one of the NO videos + gaze overlay (P5 data). This time, two of the researchers independently coded the video, with the same approach of using four phases. Then, they used MAXQDA to calculate inter-coder agreement. The agreement was >80%. After confirming the robustness of the codes, one researcher coded the rest of the data.

3.2.2 Macro-Task Classification

The goal of macro-tasks classification is to identify what participants are doing during different stages of the segmentation process (from the segmenters' perspective), and what are the domain-agnostic phases and patterns of higher-level actions during segmentation. The input data stream is a combination of participants' handwritten task-outline, task labeling, RTA and CDM. We can only apply this coding to our revised studies (P5-P10), since we did not explicitly capture task outlines in the initial study. The participants' original task-outlines (Figure 3.6 a) had to be processed both to remove domain-specific comments and because they did not always match up exactly with the observed task sequences.

Task-outlines were not always complete. Sometimes participants were doing observable tasks in the NO videos, but do not talk about them in their task-outlines. If participants have missed anything, we use the NO videos to extend and complete the task-outlines. In addition, participants may talk about lots of things that were not necessarily relevant to their actual task or useful for analysis. Therefore, we removed those

from the task-outlines as appropriate. We modified original task-outlines by:

1. **Simplifying items in task-outline:** We simplified items of the tasks into only those we want to do a more in-depth analysis on. We cleaned up the text and removed any unnecessary item. By simplifying, we only saved those items that play an important role in segmentation, and that we could associate eye-gaze data to them.

2. **Expanding items in task-outline:** We expanded the original task-outlines, using the NO labeling step (Step 4 in hybrid protocol described in Section 2.1.3). The goal was to cover all activities that were observed doing during the segmentation, even those not reflected in original task-outline.

3. **Grouping items of task-outline:** Using the expanded task-outlines we grouped items into higher-level categories or macro tasks.

Table 3.3 shows our domain-agnostic macro-task categories. Figure 3.6 b) shows modified task-outline with macro-task categories for P5.

Table 3.3: Macro-task Classification

Macro-Task	Description
Setup	Any task related to setting up, e.g., loading data, saving, and setting view
Inspection	Tasks related to inspecting ROIs, e.g., identification, looking, and scanning
Tool Usage	Using segmentation tools for marking structures (e.g., drawing) and function manipulation (e.g., changing threshold)
Review	Tasks related to reviewing segmentation results, including editing
2D Navigation and View	Moving through slices or using alternative 2D views
3D Navigation and View	Using 3D models, rendering and navigation

Coding agreement: is achieved by viewing all three reviewers' codes on a single videos for P5-P10. Inter-coder agreement was calculated using MAXQDA (agreement in percentage). Once all the coders finished a video and uploaded them to a shared drive, they used MAXQDA, to check inter-coder agreement. If the agreement was <80%, researchers reviewed codes to identify trouble spots and recode if necessary. After reaching the agreement (for P5 and P6 data), the code set was considered to be robust, and two of the researchers coded remaining participants' videos, separately (P7-P10).

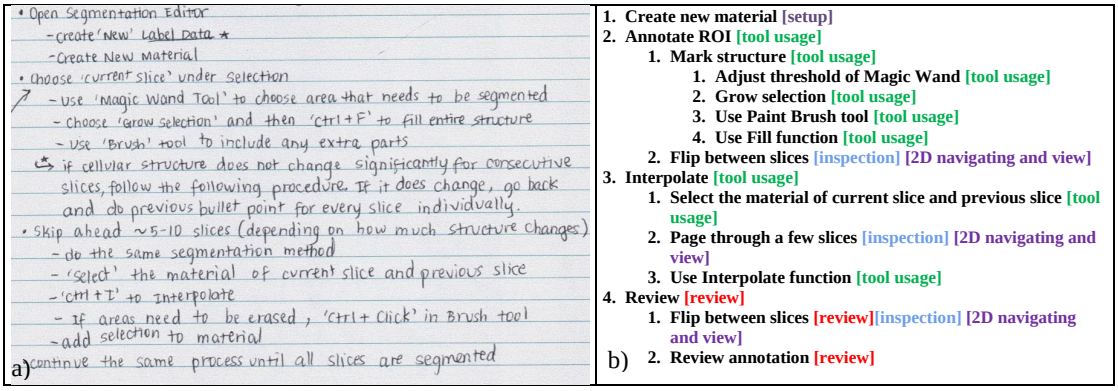


Figure 3.6: a) Original task-outline. b) Modified task-outline with macro-task categories for the original one.

3.2.3 Tool-feature/Data/Strategies Classification

The goal of Tool-feature/data/strategies classification is to identify what expert and novice segmenters state about the tool/data they work with and the strategies they employ during segmentation. This classification helped us pull all the text related to the following attributes:

- Segmenters' ideas about the segmentation process based on attributes such as accuracy, efficiency, segmentation strategies, and tool flaws.
- Tool features.
- Data characteristics.

Our data for this part includes transcriptions of CTA and RTA videos from field studies. CTA and RTA transcriptions provided us with lots of semi-structure conversations. Based the above attributes we introduce a classification that facilitated taking out all the text corresponds to what participants are stating about each attribute. We can only apply this coding to our revised studies (P5-P10), since we did not explicitly capture RTA in the initial study.

At first we identified all those parts in the videos where participants are talking about the segmentation process, strategies, dataset, and tool. We chose to transcribe word by

word to make sure we captured everything. A group of four researchers was responsible to complete the transcription task.

After transcribing all the audio/video, we used a structure method of coding to organize our data and make sense of it in a rich and revealing way. We listed possible tool features and functionalities. "Tool Feature" is any software/feature segmenters use to look at or mark their data, while "Data" is the actual image they are looking at. Please note that tool and "Tool Features" are different. When users are talking about segmentation tool, they usually mean the main software they use for segmentation (e.g. Amira a semi-automated segmentation tool). By "Tool Features" we mean any feature of the software they use during the segmentation (e.g., marking or visualization features).

We argue that all 3D image segmentation tools can be classified as one of the "Tool Features" shown in table 3.4.

Table 3.4: Segmentation "Tool Features" classification.

Tool feature		**Description**
Marking	Draw	Any tool feature for drawing contours, e.g. paint brush tool.
	Fill	Any tool feature for filling structures e.g. flood fill.
	Semi-Automated mark	Any tool to do segmentation in a semi-automated way, e.g. by placing seed points
Visualization	View 2D	Tool feature to help view image in 2D
	View 3D	Tool feature to help view image in 3D

Also, we classified the characteristics of segmentation "Data" as shown in Table 3.5. We used items of these tables as codes to organize our transcriptions (e.g. Marking is a parent code with three child codes: *Draw*, *Fill* and *Semi-Automated Mark*). We also introduce developed codes to capture: 1) Time (efficiency); 2) Accuracy of segmentation; 3) Segmentation Strategy; and 4) Tool Flaw(Missing functionality, Incorrect functionality, and Complaints.

Coding procedure and agreement: Four researchers transcribed all CTA and RTA videos for the participants. Then, they classified the data with the "Tool/Data" codes in Tables 3.4, and 3.5. Researchers used MAXQDA for transcription and coding in four passes:

1. **Pass 1:** Researcher 1 made a pass through all transcriptions. If possible he/she

Table 3.5: Classifying characteristics of "Data".

Data		**Description and example**
3D Structure		Final shape in 3D, e.g. blobby shape, sphere, has tunnels?
2D Structure	Structure complexity	How complex is the shape? Considering boundary complexity, number of holes and branches.
	Smoothness	How smooth is the boundary of a structure?
Image Quality	Gradient	Is there a directional change in the intensity or color in an image? Is there a line that marks where one region ends and the other begins?
	Texture	What are the patterns in an image? Qualities of the inside region versus the,outside
	Resolution	Are Images with enough details?

annotated each chunk of data (could be a sentence or even a paragraph) with its corresponding code from tables. *Example: If a user is talking about how complex one structure is, then the text corresponding to it in transcriptions should be coded as structure complexity.*

2. **Pass 2:** Using MAXQDA, Researcher 2 pulled all chunks for a given code and made sure that the code was actually appropriate for that chunk. If needed he/she annotated the chunk with another code or expands the chunk.

3. **Pass 3:** Researcher 3 pulled all unmarked texts and checked to make sure that the texts do not have any invalid information.

4. **Pass 4:** Researcher 4 pulled out all text again and checked if the chunks are coded appropriately.

3.3 Results

We analyzed the coded data using the three following methods:

1) Micro-task frequency analysis: Using the micro-task coding scheme, we computed code frequency and average duration of each micro code interval over the entire observation. Micro-task frequency analysis helps understand what low-level actions/tasks

participants take most often and how interleaved these actions are.

2) Macro-task frequency analysis: Using the macro-task categories, we computed code frequency and average duration of each macro-task category interval over the entire observation. Macro-task frequency analysis helps understand what higher-level tasks participants take most often and how interleaved these tasks are.

3) Snapshot analysis: We analyzed the data using video snapshots in order to distinguish how patterns change over time for different tasks. "Snapshots" are selected video segments corresponding to each of the macro-codes and task items in modified task-outlines. For this analysis, we focused on code frequencies for eye-gaze location (specifically "Boundary" vs. "Region") during each of the macro-tasks. Snapshot analysis identified what participants are doing during different stages of the segmentation process.

4) Tool-feature/data/strategies frequency analysis: Using the Tool-feature/data /strategies classifications, we computed code frequency of each category over the entire observation. This analysis helped understand what participants mostly talk about during a segmentation process and what segmentation strategies they have in mind.

3.3.1 Results of Micro-Task Frequency Analysis

We have completed micro-task action frequency analysis for all 10 participants (P1-P10). For P7, we analyzed the two observations separately and in two sessions (Session 1 and Session 2). P7 was an expert participant who segmented two different data sets in two different NO sessions.

For this analysis, we first count the frequency of each micro-task code over the entire valid observation/NO period (e.g., if participant eye-gaze is on the tool for 26 times, we count 26 intervals, and the code frequency for the "Tool" micro-code is 26). We then capture average duration by taking the average time spent in each interval (e.g., if a participant has 26 tool intervals, totaling 58 seconds, the average duration is $58/26 = 2.2$ seconds). We chose this approach because actions typically overlap so chunking up the video is not appropriate. Code frequency also lets us successfully capture micro-

task and actions with very short durations. Now we present our results for each of the micro-tasks.

Tool vs. Data Results: Figure 3.7 indicates that for each participant the frequency of looking at data is almost equal to looking at the tool, except in the case of participants of site 4 (P8-P10) who did not interact with the tool very much. The average duration spent looking at the data is greater than looking at the tool. The exception to this is P3, whose primary tool was a slider used to adjust a threshold value (no contour drawing). These results show that most participants go back and forth between the tool and data, but they spend more time gazing at data.

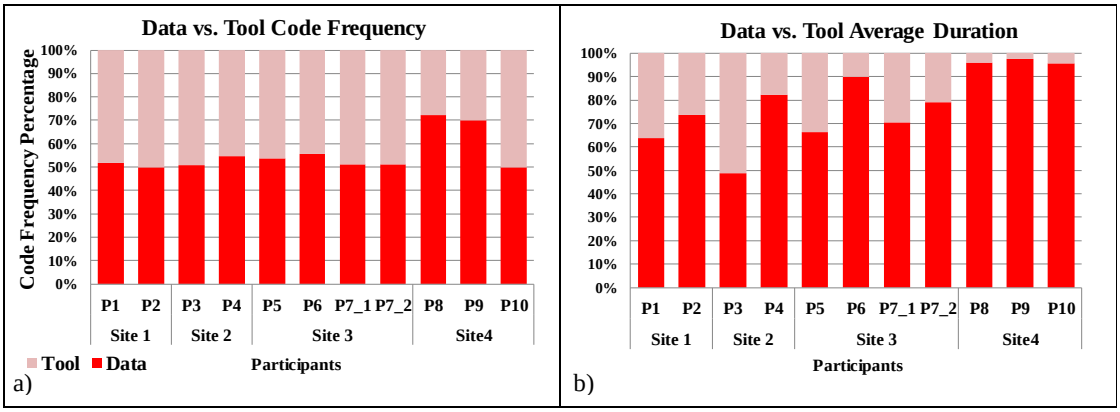


Figure 3.7: Data vs. Tool: frequency and average duration.

Boundary vs. Region Results: As shown in Figure 3.8 a, P1, P6 and P7 (Session 2) looked more at image feature boundaries while other participants looked more at regions. This was also reflected in the average duration (Figure 3.8, b). This suggests that different segmentation tools and data sets result in different gaze patterns, varying the ratio of time spent looking at regions versus boundaries. P8-P10 spent more time on non-region of interests. Since these participants segmented the structures by marking (creating circles to capture structures), they spent more time on gazing at non-region of interests (close to the regions) to make sure they placed the circle correctly (in terms of size to cover the whole structure).

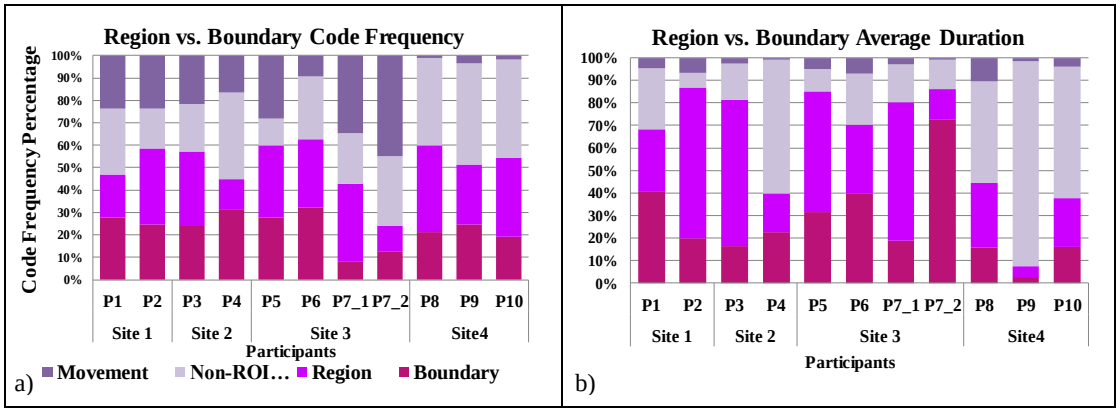


Figure 3.8: Region vs. Boundary: frequency and average duration.

Navigation Results: The participants showed no particular patterns for overall navigation (Figure 3.9). However, most participants (except P3) both paged more frequently — and spent more time paging— than any other navigation type.

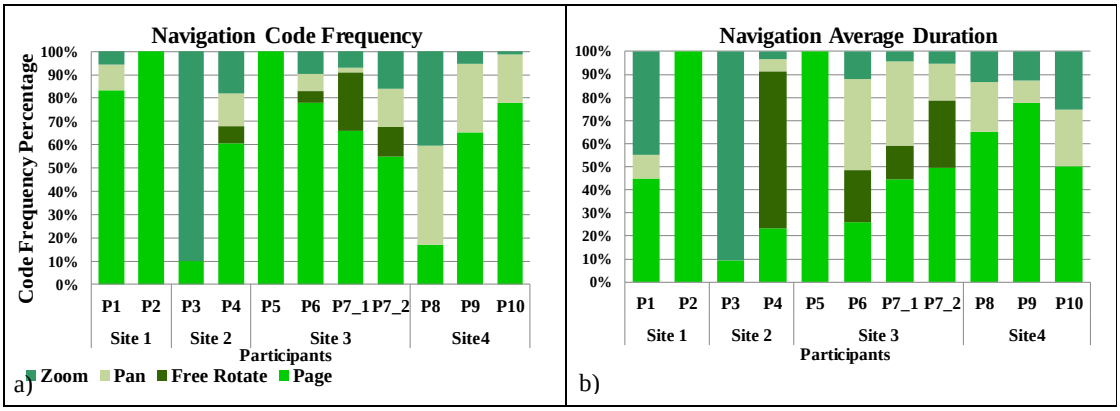


Figure 3.9: Navigation: frequency and average duration.

Marking Results: Regarding "Marking" micro-tasks, as shown in Figure 3.10, those participants who have higher code frequency for "Boundary" than "Region" also have higher code frequency for "Draw" compared to "Fill", and vice versa. P8-P10 spent more time on non-region of interests. Since these participants segmented the structures by marking (creating circles to capture the structures), they spent more time on gazing

at non-region of interests (close to the regions). This shows a potential relation between different gaze patterns and Marking actions. Those who have a boundary-based gaze pattern tend to do the segmentation by drawing while those with a region-based gaze pattern complete the segmentation process by filling the structure. Participants who segments the structure by marking (e.g., putting circles) tend to look more at non-ROIs. P3 did not take any marking action during the observation since they were just adjusting a slider.

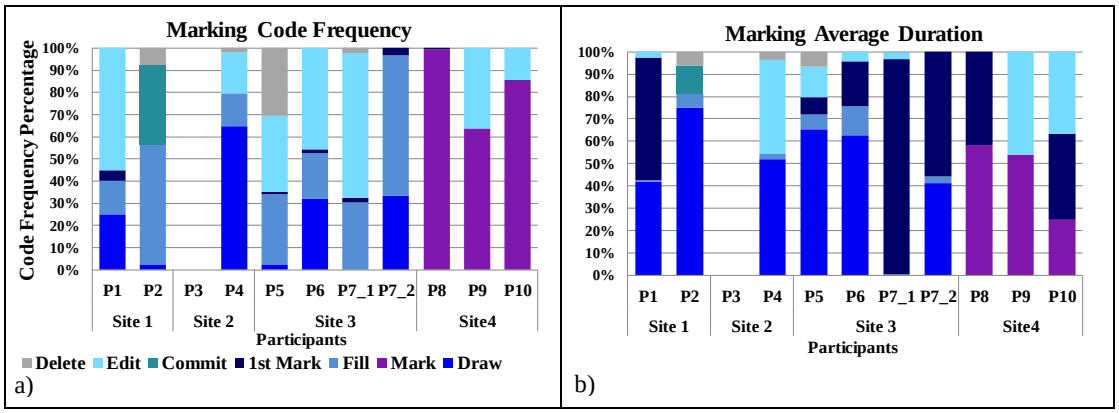


Figure 3.10: Marking: frequency and average duration.

Review Results: According to Figure 3.11 there are code similarities for P1 and P2 of site 1, P4 of site 2, P5 and P7 (Session 1) of site 3, and P8-P10 of site 4. P3 has a completely different code pattern (P3 was adjusting a slider). P7 (Session 2) has a higher code frequency for Access data. Both P5 and P6 spend more time on "Mark Review" comparing to P7.

3.3.2 Macro-Task Frequency Analysis and Results

Using macro-task categories, we completed macro-task frequency analysis for 6 participants (P5-P10) whom we observed in the revised studies. For P7, we analyzed the two observations separately. P7 was an expert participant who segmented two different data sets in two different NO sessions. Similar to micro-task analysis, we first count the frequency of each macro-task code over the entire valid observation/NO period. We then

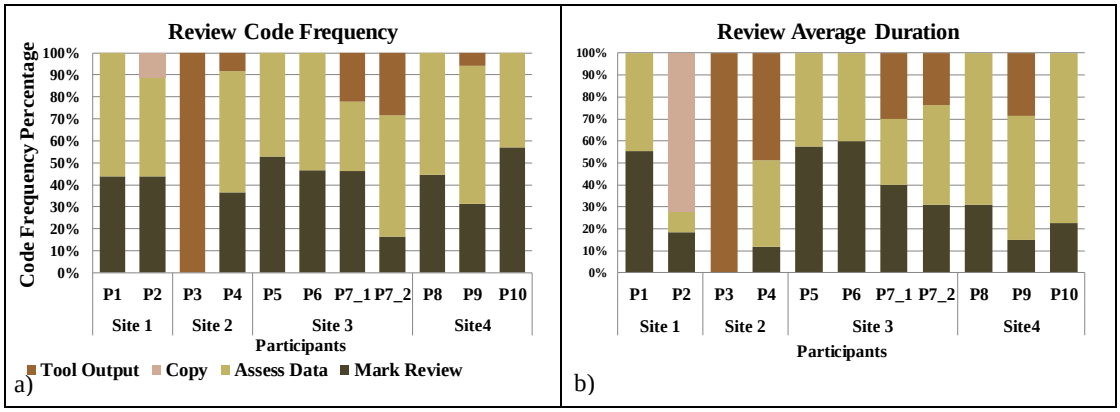


Figure 3.11: Review: frequency and average duration.

capture duration by taking the average time spent in each interval (e.g. if a participant has 20 "Review" task intervals, totaling 60 seconds, the average duration for "Review" macro-task is $60/20 = 3$ seconds).

Figure 3.12 shows the category frequency and average duration for each of the participants. Different participants have different patterns of task frequency and average duration. The "Tool-usage" category has the highest frequency which indicates that participants mostly work with the tools and use different mechanisms (e.g., drawing) to segment the structures. Although "Setup" frequency is low for all participants, the average duration indicates that "Setup" is a time intensive process for some of the participants (P8 and P10). One interesting difference between experts and novices is that the experts spent much more time viewing/working with the 3D structure (e.g., P7 and P10 are both experts and spend more time on 3D views/navigation).

3.3.3 Snapshot Analysis and Results

In this section, we analyze the data using video snapshots in order to distinguish how patterns change over time for different tasks. We used tasks based on the participant's task-outline. For task snapshot analysis, we only focused on four micro-task eye-gaze location codes: "Boundary", "Region", "Non-ROI feature" and "Movement". We have snapshot analysis for the 6 participants whom we observed in our revised studies (P5-P10). We have two types of snapshot analysis:

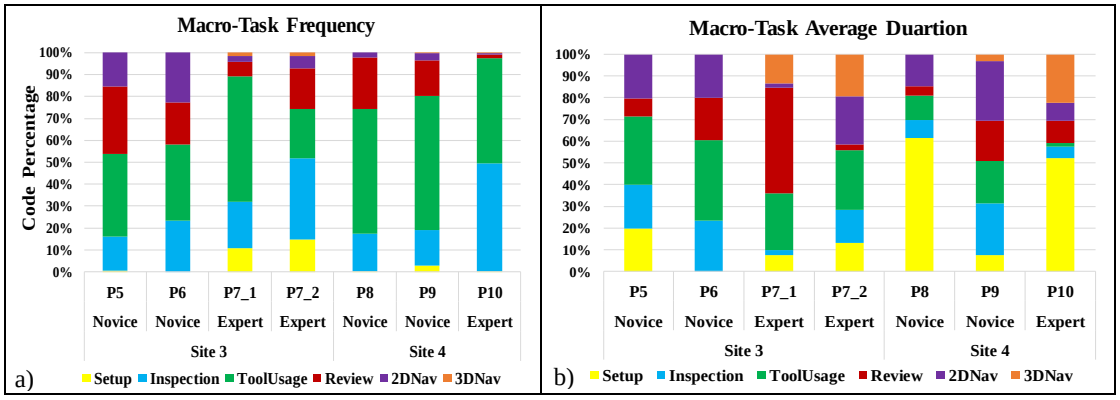


Figure 3.12: Frequency and average duration for each macro-task.

1. Macro-task snapshot analysis: Count micro-task eye-gaze codes (Boundary, Region, Non-ROI feature and Movement) frequency and average duration for each of the macro-task categories. We have macro-task snapshot analysis for 6 participants whom we observed in our revised studies (P5-P10).

2. Task-outline snapshot analysis: Count micro-task eye-gaze codes (Boundary, Region, Non-ROI feature and Movement) frequency and average duration for selected task items of the modified task-outlines. This analysis helped us to determine what participants are doing during different segmentation tasks

3.3.3.1 Macro-Task Snapshot Analysis

We analyzed each macro-task category (e.g., Setup) in detail by assigning one snapshot to each macro-task and then count the eye-gaze micro-task frequency (Boundary, Region, Non-ROI feature and Movement) within each snapshot (1 count being 0.1 seconds). The final code percentage for each macro-task is separately computed by summing up all frequencies of the code in each snapshot. Figure 3.13 shows macro-task snapshot results (frequency and average duration) for site 3 (P5-P7) and site 4 (P8-P10). Please note we observed P7 in two sessions.

As shown in Figure 3.13 a) and b), similar to micro-task analysis, P5 and P7 (Session 1) look more at regions and also spend more time looking at regions for almost all the

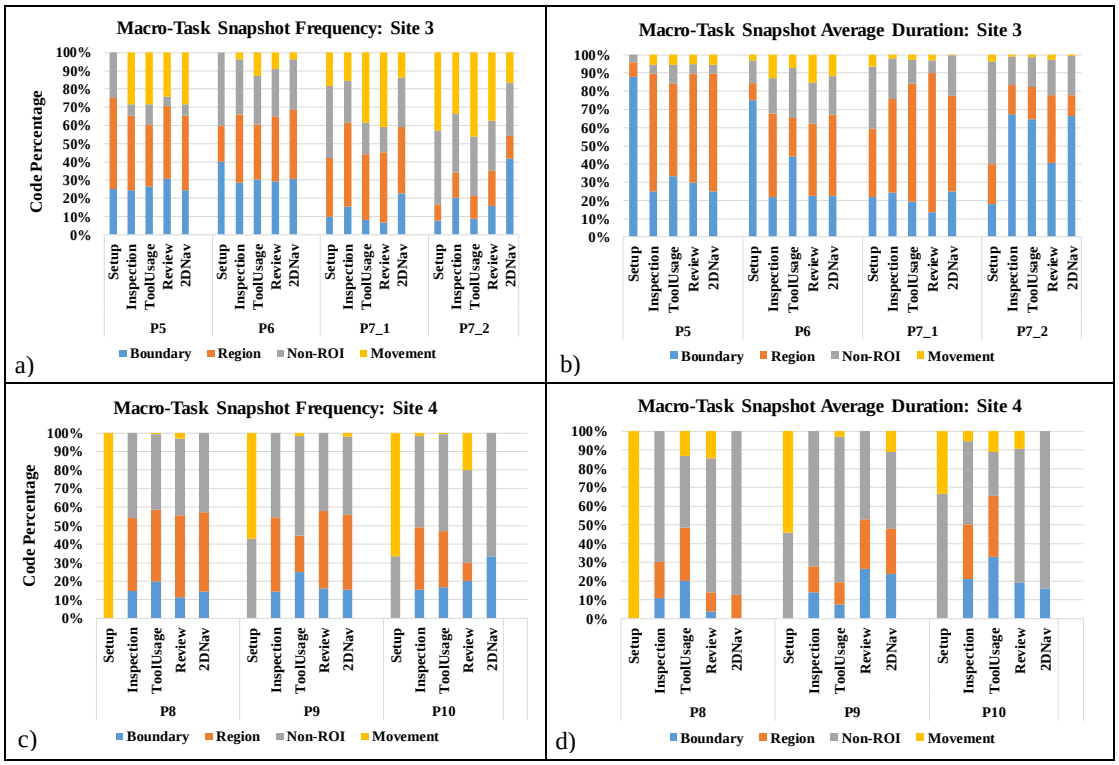


Figure 3.13: Macro-task snapshot analysis: Frequency and average duration for participants: a) and b) Site 3 (P5-P7); c) and d) Site 4 (P8-P10).

macro-tasks (expect "Setup" for P5 which the participant spends more time looking at the boundary). In our micro-task analysis, we showed that these participants use the "Fill" approach to do the segmentation which involves looking more at regions. On the other hand, P6 and P7 (Session 2), who use paint brush tool for segmentation, have higher code frequencies for boundary compared to region, and spend more time looking at boundaries. Comparing to novice users (P5 and P6), P7 (expert user) has a different pattern in terms of "Movement". Both P5 and P6 have fewer movements. At a high level, the expert paged back and forth between slices multiple times between fills/drawings, whereas the novice just filled or drew. One other clear difference between the expert and the novices is that the expert spent much more time up front evaluating the data as a whole before beginning marking.

Comparing to Site 3, participants of Site 4 have different patterns for frequency and

average duration. Figure 3.13 c) and d) shows that during each of the macro-tasks participants of Site 4, who segment the structures by marking (e.g., putting circles) tend to look more at non-ROIs. We could not identify much difference between the expert and the novices of Site 4.

3.3.3.2 Task-Outline Snapshot Analysis

In the previous section, we completed snapshot analysis for each of the macro-tasks (e.g., how frequently does a participant look at boundaries during the "Review" macro-task). But we are also interested to know what participants are doing during different tasks (please note by task we mean a task item included by the participant in the modified task-outlines). Our goal is to identify if there is any pattern between different/similar tasks for different participants, and how similar the gaze patterns are for repeated instances of one task example.

For task-outline snapshot analysis we picked two participants from each site (one expert and one novice). We also picked two different individual tasks for each of the participants. Table 3.6 summarizes the participants and tasks for task-outline snapshot analysis. We chose the task examples in a way to insure similarities in task types. For instance, for P6 and P7, task examples 1 are similar because both tasks include tracing boundaries and using a drawing (brush/lasso) tool. Task examples 2 are also similar as they both include flipping between slices.

Table 3.6: Participants' Task Examples

Participant	**Status**	**Task Example 1**	**Task Example 2**
P6	Novice	Use Paint Brush Tool	Flip Between Slices
P7_Session 2	Expert	Trace ROI boundary and use Lasso tool	Flip between slices
P8	Novice	Identify Cell	Drop blue disks to fill cell
P10	Expert	Identify Cell	Drop blue disks to fill cell

We analyzed tasks in detail by assigning one snapshot to each task example and then count the code frequency within each snapshot (similar to macro-task analysis). Again, we only focused on four micro-task eye-gaze location codes (Boundary, Region, Non-ROI feature and Movement).

Results for P6 and P7 (Session 2): For P6 who is a novice participant at Site 3, we chose the two individual sample tasks "Example 1: Use Paint Brush Tool" and "Example 2: Flip Between Slices". We captured 9 repeated task instances of P6 task example 1 (T1-T9), and 8 instances of task example 2 (T1-T8). For the expert segmenter (P7 Session 2), we chose the two individual sample tasks "Example 1: Trace ROI boundary and use Lasso tool", and "Example 2: Flip Between Slices". We captured 6 repeated task instances of task example 1 (T1-T6), and 7 instances of task example 2 (T1-T7).

Figures 3.14 shows the analysis results for P6 and P7 Task example 1. P6 uses the paint brush tool for annotation to capture the boundaries. P7 uses the lasso tool to trace the boundaries. Our previous results showed that participants who use drawing for segmentation, also look more at boundaries (boundary-based approach). Therefore, we expect P6 and P7 (session 2) to look more at boundaries while doing the annotations. Results of analysis confirm this. Both P6 and P7 look more at boundaries and spend more time on boundaries during task example 1.

Figures 3.15 shows the analysis results for P6 and P7 Task example 2. Although the task is similar for both participants (Flip between slices), P6 and P7 have different gaze patterns for the task. The duration for "Movement" is higher for P6. P7 still spends more time looking at boundaries and has lower average duration for movement. Both participants have higher code frequency and average duration for "Non-ROI" code in their task example 2 (compared to task example 1).

P6 and P7 (Session 2) have a boundary-based strategy of segmentation and look/spend more time on boundaries. However, they have different task-outlines and gaze patterns even if they worked with similar tools and data sets. In addition, gaze patterns were different in different individual tasks, but repeated instances of tasks show more similar patterns.

Results for P8 and P10: For P8 (novice) and P10 (expert), we chose the two individual sample tasks "Example 1: Identify Cell" and "Example 2: Drop blue disks to fill cell". For both participants, we captured 10 repeated instances of task example 1 (T1-T10). For task example 2, P8 has 9 instances (T1-T9), and P10 has 12 instances (T1-T12). Figures 3.16 shows the analysis results for P8 and P10 Task example 1. They both look more at non-ROI features. However, P8 almost does not look at boundaries, while P looks more and spends more time on boundaries.

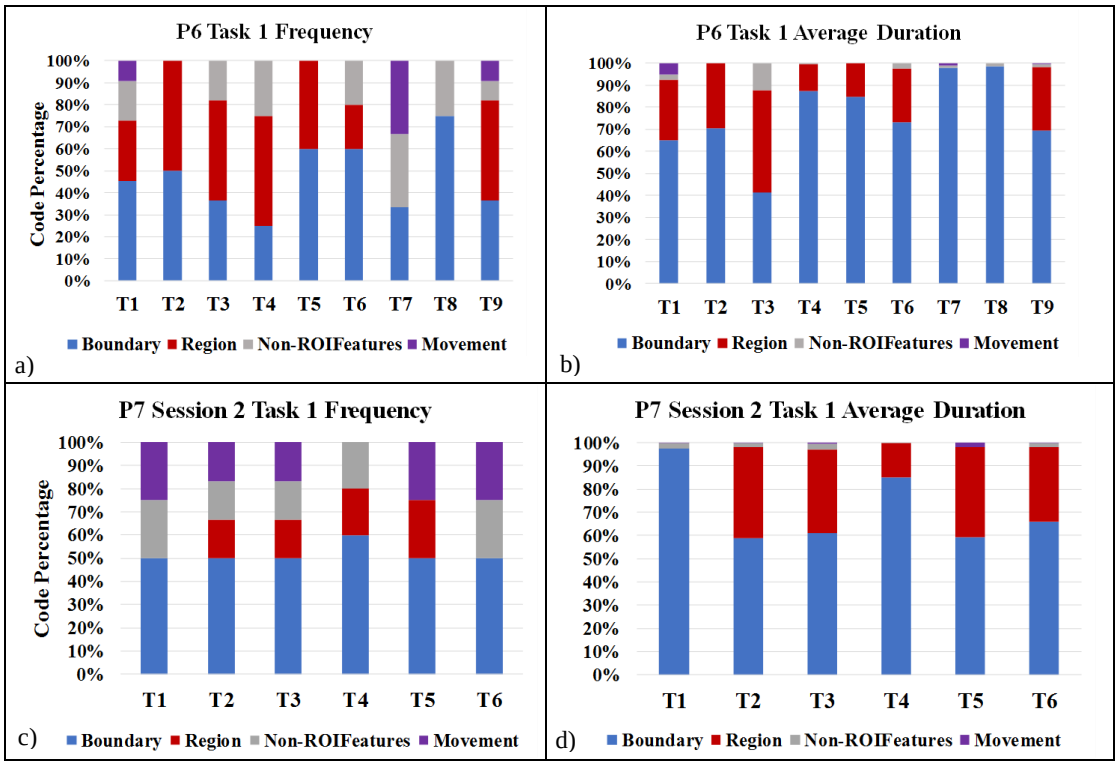


Figure 3.14: Code frequency and average duration for Task Example 1. a) and b) P6, Use Paint Brush Tool; c) and d) P7, Trace boundary and use lasso tool.

Figure 3.17 shows the analysis results for P8 and P10 Task example 2. Both participants use marking mechanisms (dropping circles) for annotation. Although the task is similar for both participants, but they have different gaze patterns for the task. While P10 almost does not look at regions, P8 looks more and spends more time on boundaries. Both participants have higher frequency for "Non-ROI" code.

Again, as we can see in figures, P8 and P10 have non-region based strategy of segmentation and look/spend more time on "Non-ROIs". They have different task-outlines and gaze patterns even if they worked with similar tools and data sets and cover similar tasks. In addition, gaze patterns were different in different individual tasks, but repeated instances of tasks show more similar patterns for each participant.

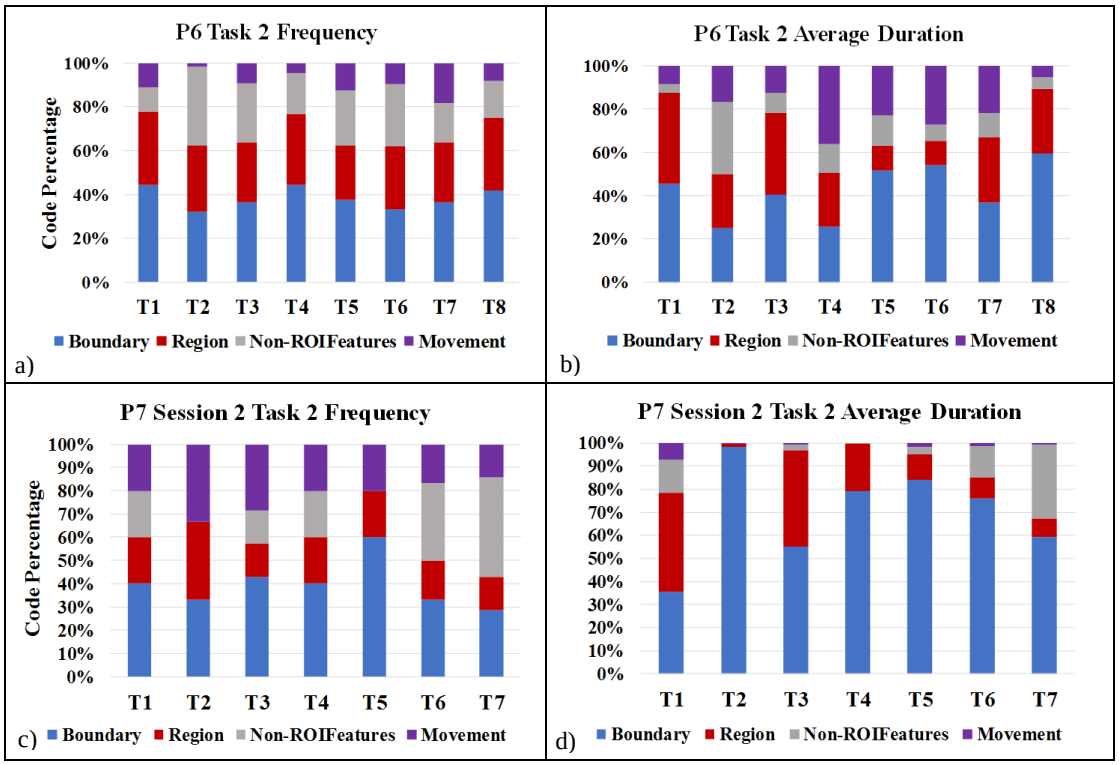


Figure 3.15: Code frequency and average duration for Task Example 2 (Flip Between Slices). a) and b) P6; c) and d) P7

3.3.4 Results of Tool-feature/Data/Strategies Frequency Analysis

We counted the code frequency for these categories: *Measures* (efficiency and accuracy), *Strategy*, *Tool* (marking and visualization), *Data* (2D and 3D structure), and *Tool Flaw*. Table 3.7 and 3.8 show the code frequencies and percentages. P5, P6, P8, and P9 are novice participants while P7, P10, are experts.

The most frequent code category is *Strategy* which captures all parts of a transcription where participants were talking about segmentation strategies. Experts in general talked more about 3D structures. Analysis from NO videos also showed they spend more time working with 3D structure/views.

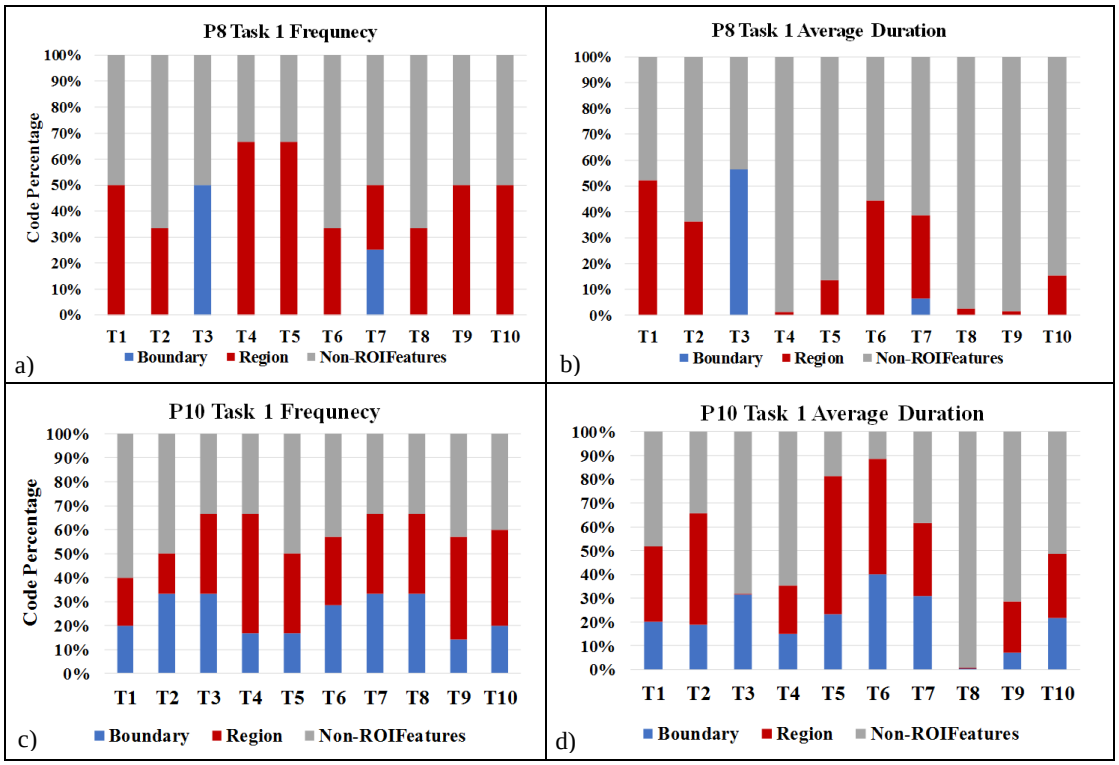


Figure 3.16: Code frequency and average duration for Task Example 1, Identify Cell. a) and b) P8; c) and d) P10

3.4 Summary of Results and Discussion

RQ1: Using our micro-task coding scheme and macro-task classification, we could successfully quantify where segmenters look, and what their low-level action and higher-level tasks are. We identified that depending on the segmentation tool and the data set, participants utilized one of these three segmentation strategies: Region-based (seeing data as regions to be filled/ captured), boundary-based (seeing data as boundaries to be demarcated by drawing contours), and non-ROI-based. Additionally, if participants explicitly captured a structure by drawing, they looked more at the boundaries, filling resulted in more gazing at regions., and marking (e.g., putting circles) involved looking more at non-ROIs. Table 3.9 shows participants' segmentation approaches.

For region and boundary based strategies, participants exhibited approximately equal

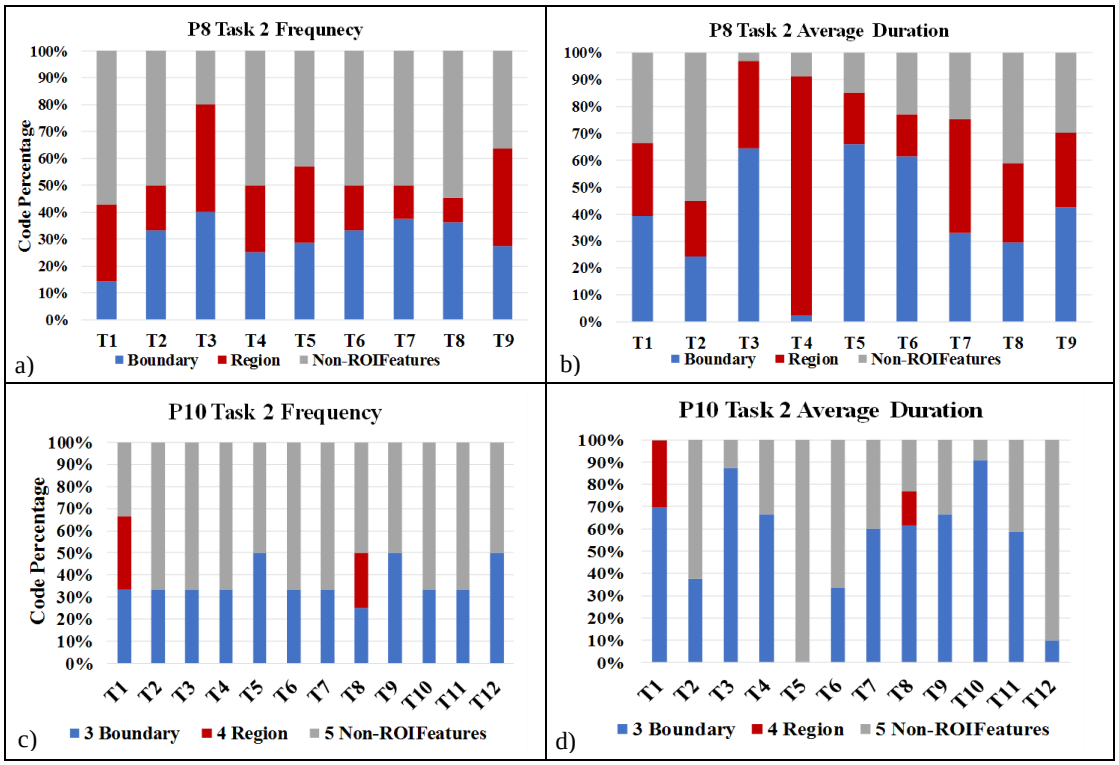


Figure 3.17: Code frequency and average duration for Task Example 2, Drop blue disks to fill cell. a) and b) P8; c) and d) P10.

amounts of gaze movement between the data and the tool. For non-ROI based strategies, participant looked more at data. All participants, on average, spent more time gazing at data than tools.

RQ2: Different participants had different task-outlines and gaze patterns even if they worked with similar tools and data sets. One participant utilized two different segmentation approaches (same data set and tools, but a different structure). In addition, gaze patterns were different in individual task examples, but repeated instance of the same task had similar patterns.

We can see a relationship between tool categories frequency and the segmentation strategies we captured using the micro-task analysis. For example, if a participant has a region-based strategy and performs segmentation by filling the structure, he/she talks

Table 3.7: P5-P7 code category frequency for Measures, Strategy, Tool, Data, and Tool Flaw

Code Category			P5:Novice		P6:Novice		P7:Expert	
			Freq.	Pct. %	Freq.	Pct. %	Freq.	Pct. %
Measure	Time		2	3.0	4	4.2	4	4.2
	Accuracy		2	3.0	11	11.5	18	18.8
Strategy			9	13.6	23	24.0	17	17.7
Tool	Mark	Draw	9	13.6	4	4.2	6	6.3
		Fill	13	19.7	1	1.0	5	5.2
		Semi	0	0.0	6	6.3	5	5.2
	Vis	View 2D	5	7.6	4	4.2	9	9.4
		View 3D	6	9.1	5	5.2	8	8.3
Data	3D Strcut.	3D strcut	5	7.6	5	5.2	8	8.3
	2D Struct.	Complexity	3	4.5	4	4.2	1	1.0
		Smoothness	0	0.0	2	2.1	0	0.0
	Quality	Gradient	3	4.5	4	4.2	4	4.2
		Texture homogenity	6	9.1	4	4.2	2	2.1
		Resolution	0	0.0	3	3.1	0	0.0
Tool Flaw		Missing functionality	1	1.5	2	2.1	0	0.0
		Incorrect functionality	2	3.0	4	4.2	2	2.1
		Complaints	0	0.0	10	10.4	7	7.3

Table 3.8: P8-P10 code category frequency for Measures, Strategy, Tool, Data, and Tool Flaw

Code Category			P8:Novice		P9:Novice		P10:Expert	
			Freq.	Pct. %	Freq.	Pct. %	Freq.	Pct. %
Measure	Time		3	4.6	4	6.5	9	14.5
	Accuracy		11	16.9	10	16.1	5	8.1
Strategy			11	16.9	9	14.5	18	29.0
Tool	Mark	Draw	0	0.0	0	0.0	0	0.0
		Fill	0	0.0	0	0.0	0	0.0
		Semi	9	13.8	6	9.7	6	9.7
	Vis	View 2D	0	0.0	3	4.8	2	3.2
		View 3D	7	10.8	4	6.5	6	9.7
Data	3D Strcut.	3D strcut	0	0.0	0	0.0	4	6.5
	2D Struct.	Complexity	2	3.1	3	4.8	4	6.5
		Smoothness	0	0.0	1	1.6	0	0.0
	Quality	Gradient	9	13.8	12	19.4	1	1.6
		Texture homogenity	5	7.7	6	9.7	1	1.6
		Resolution	1	1.5	2	3.2	0	0.0
Tool Flaw		Missing functionality	4	6.2	0	0.0	2	3.2
		Incorrect functionality	2	3.1	1	1.6	3	4.8
		Complaints	1	1.5	1	1.6	1	1.6

Table 3.9: Participants' Marking methods and segmentation approaches

Participant	Marking Method	Segmentation Approach
P1	Draw Contours	Boundary-based
P2	Fill Structure	Region-based
P3	Adjust Slider	Region-based
P4	Draw Contours	Boundary-based
P5	Fill Structure	Region-based
P6	Draw	Boundary-based
P7 (Session 1)	Fill Structure	Region-based
P7 (Session 2)	Draw	Boundary-based
P8	Put Circle Marks	Non-ROI-based
P9	Put Circle Marks	Non-ROI-based
P10	Put Circle Marks	Non-ROI-based

more about the *Fill* tool in the transcriptions. This is true for the boundary-based strategy and the *Draw* tool code.

Results from both macro-task and tool-feature/data/strategy analysis show comparing to novices, experts spent more time working with 3D structures/views and they talked more about 3D. In general experts showed more interest in working with 3D structures/views. Here are a few transcription examples of our participants talking about 3D structures/views (experts supported working with 3D structures/views to better understand 2D slices; novices are reluctant to work with 3D structures/views).

- **P5 (Novice)- Showing lack of interest to work with 3D structures/views for understanding 2D slices:**
  - *"I do not often look at the 3D shape."*
  - *"No, I don't go this detailed (for making judgments) with the 3D structure."*
- **P6 (Novice)- finding it difficult to work with 3D structures/views:**
  - *"Im not sure how to get to the 3D rendering.*
  - *"I very rarely look at 3D side view, just because they are somewhat confusing".*
  - *"3D views make it very difficult to understand what you are looking at and often its because there were jumps in the data".*

- **P7 (Expert)- Effective use of 3D view:** *"I think for this particular data set there is high retention of detail and it is quite useful to see this in **3D**, from different orthogonal views, and to see what that looks like from a different angle."*

- **P8 (Novice)- Difficulty working with 3D structures/view:** *"I think its very difficult [working with 3D for the current 2D slice], even as a new person. Theres a lot of discrepancies between cell boundaries. Theres sometimes, like even for this for example, theres a tear through this slice and I might not know exactly how large I should make certain annotations."*

- **P10 (Expert)- Understanding of 2D slice and 3D structures:** *"So were not tracing any of the intermediate stuff because we know that its a cylinder that extends so weve got the sense of the 3D geometry that were tracing and so we dont have to draw everything around it everywhere."*

To evaluate this result from another perspective, we computed all the code frequencies related to working with 3D structures/3D views, 3D navigation, and mapping 2D slices to 3D structures for P5-P10. Results indicate that experts worked more with 3D structures/views mainly to:

- Identify the correct 2D contours and segment them accurately.

- Double checking the result of the segmentation for different 2D slices.

- Evaluate the result of current segmentation (3D structure) from different views to capture the necessary details to improve the result of necessary.

Figure 3.18 shows the frequency of working with 2D slices based on 3D structures/views for both experts and novices. As explained, experts worked more with 3D view/structures to correctly identify the 2D contours for the manual segmentation.

Based on these study results, we argue that given a 3D structure and slicing plane, experts can: 1) predict the 2D contour; 2) predict how 2D contour changes with small view changes; and 3) identify invalid 2D contours.

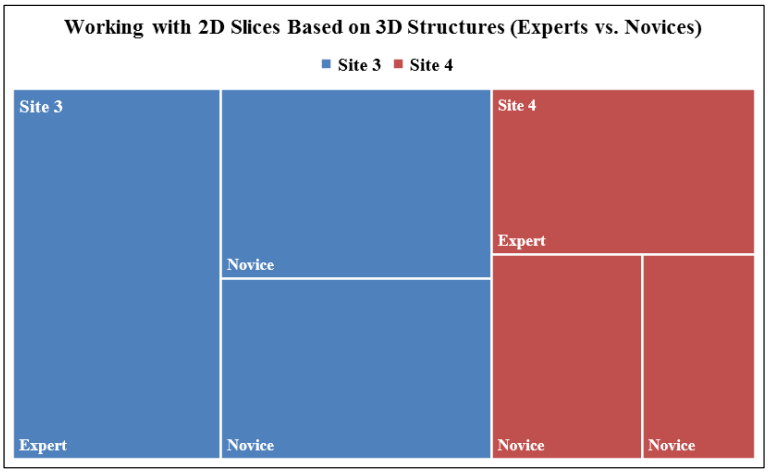


Figure 3.18: Working with 2D Slices Based on 3D Structures (Experts vs. Novices).

3.5 Conclusions

We presented a novel taxonomy for coding observations of 3D segmentation process, and used the coding scheme to successfully capture and analyze 10 segmenters' perceptual cues and low-level actions (micro-tasks). We could identify where experts look and what their low-level (micro) and higher-level (macro) segmentation actions and tasks are. We used the coded data to identify three potential segmentation strategies (region, boundary, and non-ROI-based). We also introduced our macro-task classification mechanism and observed how gaze and task patterns change over time for different segmentation strategies. In addition, we compared novices with experts in terms of both micro and macro-tasks to capture their differences. For example, one key difference between experts and novices is that the experts spend much more time viewing/working with the 3D structure. They also, comparing to novices, have better understanding of 2D slices of 3D structures.

To ensure reliable study results with this small sample size, we focused on rigorous data collection from multiple sources (video, audio, eye-gaze) and cross-validation between the different data streams. Our focus was not on statistical sample analysis but rather on a repeatable coding scheme and task-flow visualization to identify particular patterns both within and between participants. With our analysis, we effectively captured, in a quantitative manner, the tasks and behaviors observed qualitatively.

We are using these results in the next parts of our research, with the ultimate goal of creating training guidelines to improve segmentation process. As described in this chapter, the ability to visualize and infer cross-sections of biological structures is a fundamental skill for segmenters. Targeting mainly novices, we plan to develop a domain-agnostic training tool to enhance the important spatial skill of understanding 2D cross-sections of 3D structures [79].

Chapter 4: A 3D Spatial Ability Test Instrument for 3D Volume Segmentation

Manual 3D volume segmentation is completed (or evaluated) on 2D cross-sections of the 3D data and the segmenter must mentally integrate these into a coherent 3D structure. Therefore, the ability to visualize and infer cross-sections of biological structures is a fundamental skill for segmenters.

Unfortunately, there is no agreed-upon formal methodology to evaluate segmenters' 3D spatial ability independent of the overall task. In the previous part of our research, we conducted in-depth field studies with human experts, while capturing video and eye-tracking data [81, 102]. Preliminary results from these studies indicate that experts have a better understanding of the expected 3D structures and tend to use 3D views more frequently than novices in order to review segmentation results.

The ability to understand 2D cross-sections of 3D structures is an important spatial skill that is not only necessary in 3D image segmentation (e.g., using medical images such as MRIs [40]), but also contributes to performance in other science fields including engineering [36, 56, 24], geology [49, 66], biology [78, 76], and geometry [11, 70].

We hypothesize that training novices to identify valid 2D cross-sections of 3D structures (and vice-versa) would improve their overall ability to segment. The aim of this part of our research is to develop and validate a 2D cross-section testing instrument that can be used to evaluate the effectiveness of training strategies for 3D segmentation.

We begin with an existing spatial testing instrument, the Santa Barbara Solids Test (SBS for short). This test is "a psychometric measure of cross-sectioning ability with 30-item multiple choice question items to examine sources of difficulty in inferring cross-sections" [20, 21, 22].

Unlike another mental cutting test (the "Schnitte" test which measures cross-section understanding skill in participants with extremely high spatial ability [73]), the SBS test was designed to measure performance differences in participants with a normal distribution of spatial skill [21]. The measure investigates the cognitive components of cross-section understanding. It consists of 2D stimuli (images) of 3D geometric shapes

with cutting planes, where the participants pick the correct 2D cross-section. Motivated by research which emphasizes that most basic recognizable 3D forms are primitive solids [9, 67], the SBS test uses primitive geometric solids, and compound shapes made up of these solids. It has specific foils for the incorrect answers, and categorizes the questions based on number and type of primitive objects, how they are combined (touching or embedded), and the slicing plane location relative to the object (axis-aligned, oblique). The SBS test has been validated with a wide variety of participants, and in general is correlated with spatial ability as measured by the paper-folding, Vandenberg Mental Rotation tests [100], and Visualization of Views [27].

The SBS test only uses simple geometric solids and is designed to examine primitive 3D shape inference abilities. In a segmentation process, segmenters work with more complex organic/biological shapes. Therefore, we need to use a test that has more biological 3D models and resembles more complex 3D spatial tasks. For this reason, we modified the SBS test in three ways:

- First, in the segmentation task users typically have the ability to visualize the surface and data in 3D, so we moved from 2D stimuli to 3D.
- Second, we wish to calibrate how well participants perform on organic shapes in relationship to geometric ones. We also added two common organic structures, branching and holes, that were not presented in the original SBS test.
- Third, we investigated two other categories of questions that capture the segmentation process of placing a plane and drawing on sequences of planes (refer to Figure 4.1).

We have two distinct implementations for our 3D spatial ability test instrument. The "2D Cross-section Understanding Test Version 1" has a mixed combination of objects from both the SBS test (3D primitive objects) and our proposed organic shapes (e.g., branching, and a potato with a hole). This enables comparison to the SBS test. Based on this test instrument, we then created the second version of the cross-section understanding test. The "2D Cross-section Understanding Test Version 2" has two main aspects:

1. All the questions use organic/biological objects.

2. The questions cover a range of difficulty based on our novel difficulty category (see Section 4.2 for more details).

To evaluate and calibrate these tests we conducted user studies primarily using Mechanical Turks. In the following sections we describe each of these test instruments and their corresponding user studies.

4.1 2D Cross-section Understanding Test Version 1

In design, implementation, and evaluation of this test instrument, we investigate the following research questions:

- **RQ1)** How does modifying the stimuli to make it 3D change participants' responses?
- **RQ2)** What are the sources of difficulty for participants in inferring cross-sections for different 3D structures?
- **RQ3)** Is a video tutorial on 2D cross-sections more helpful than static/written instructions?
- **RQ4)** Do gender differences effect performance on understanding 2D cross-sections?

4.1.1 Test Design and Implementation

Our 3D spatial ability test instrument is implemented in Qualtrics [74] (see Appendix item "2D Cross-section Understating Test Version 1"). It has 44 multiple choice main questions distributed in three categories (Figure 4.1):

- Category 1: Given a 3D structure and slicing plane, identify the correct 2D cross-section contour.
- Category 2: Given a 2D cross-section for a 3D structure, identify the slicing plane that generated the contour.
- Category 3: Given a 3D structure and multiple slicing planes, identify the valid contour sequence that corresponds to the slices.

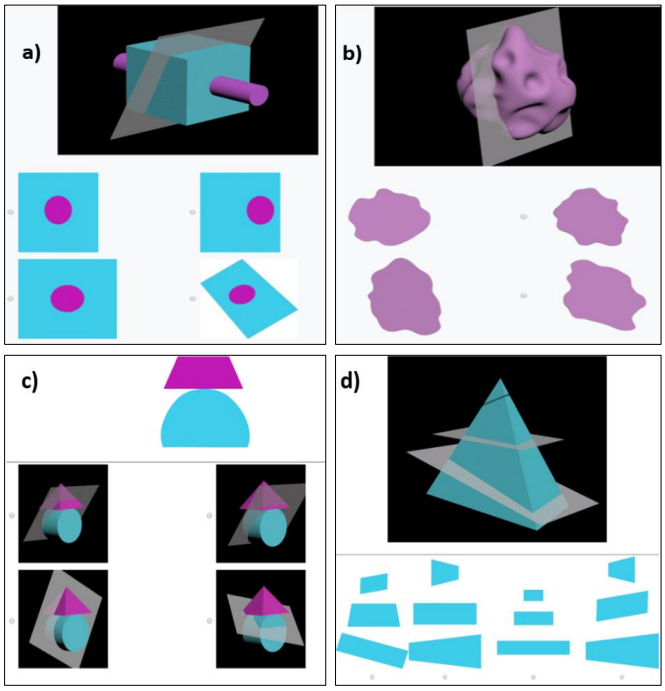


Figure 4.1: Sample test questions. a) Category 1: Embedded structure with an oblique plane; b) Category 1: Biological structure with an oblique plane; c) Category 2: Cross-section of a Joined structure; and d) Category 3: Simple object with three slicing planes.

Question items are of four different types of 3D structures: Simple, Joined, Embedded, and Biological; and two main types of slicing planes: Orthogonal (Vertical or Horizontal) and Oblique. Simple structures consist of only one object, while Joined and Embedded structures are composed of two different objects that touch or are embedded within each other, respectively. Biological shapes are more complex objects (e.g., protrusions, branches, and structures with hole) with irregular surface features. Figure 4.2 shows examples of each type of test structure and each slicing plane. Table 4.1 shows the distribution of questions based on category, object type, and plane.

One of the main differences between our test and the SBS test is the 3D stimuli for the 3D objects plus slicing plane. Instead of seeing a static image of a 3D structure with a slicing plane, the participant first sees a paused image, then the 3D structure rotates (around y axis). This rotation is added to help participants see the objects in a more

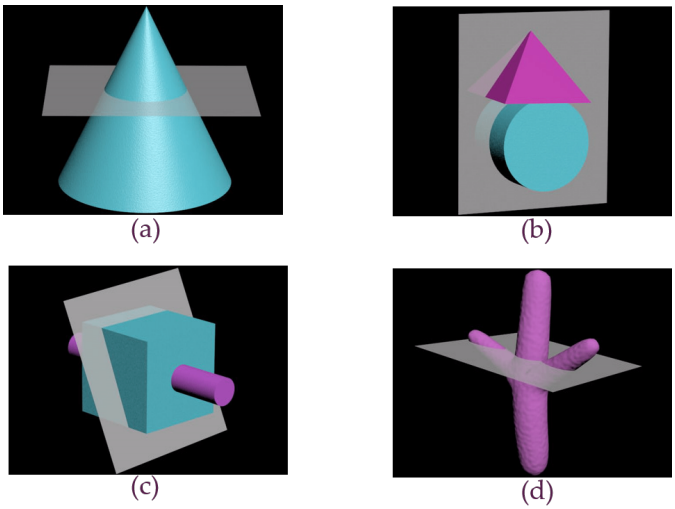


Figure 4.2: a) Simple orthogonal; b) Joined orthogonal; c) Embedded oblique; and d) Biological oblique sample objects.

Table 4.1: Distribution of the test questions based on question category, object type, and plane orientation.

	Simple		**Joined**		**Embedded**		**Biological**	
	Orth.	**Obl.**	**Orth.**	**Obl.**	**Orth.**	**Obl.**	**Orth.**	**Obl.**
Category1	5	5	5	5	5	5	1	4
Category2	0	0	1	2	0	1	0	1
Category3	0	1	0	1	0	1	0	1
Total	5	6	6	8	5	7	1	6
	11		**14**		**12**		**7**	

realistic 3D space, and therefore remove the mental step of visualizing a 3D structure from a 2D image. Test 3D structures, animations, and answer choices were created in 3D Studio Max [6], and ZBrush [71]. For the SBS questions we began with the same viewpoint as the static 2D image. For the biological shapes we mostly began with a viewpoint that put the slicing plane perpendicular to the view vector.

Similar to SBS test, each of our test question shows a main figure (3D structure with slicing plane for Category 1 and Category 3, and 2D cross-section for Category 2

questions) and four answer choices (see Figurer 4.1). As defined by the SBS test [20], each item has three types of wrong answers: 1) Alternate wrong answer (Figure 4.3 a) shows another possible slice of the test figure; 2) Combination wrong answer (Figure 4.3 b) merges two possible sections of the test figure into a hybrid shape; 3) Egocentric wrong answer (Figure 4.3 c) represents a shape that participants might visualize if they failed to change their view perspective relative to the slicing plane of the criterion figure.

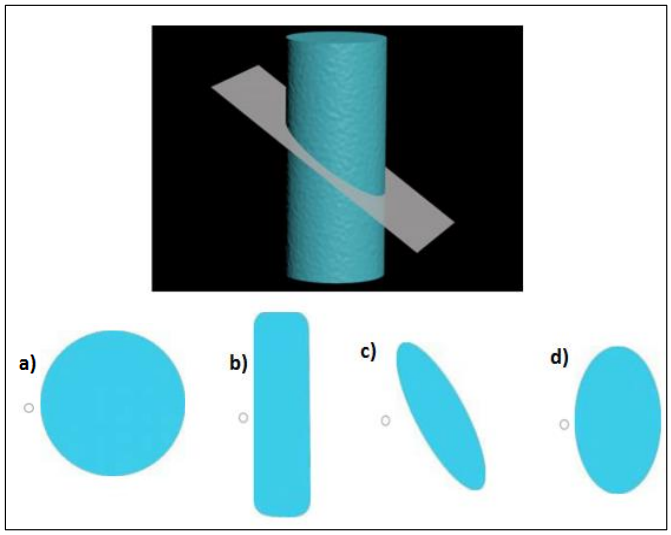


Figure 4.3: A Simple oblique test item (cylinder) and the four answer choices: a) Alternative; b) Combination; c) Egocentric; and d) Correct answer.

To help participants become familiar with the concept of cross-section, we created two versions of the guidelines. The first version is a video tutorial, and the second version is simple written instructions with an image. The written instruction defines the term cross-section with a sample (a sphere with a horizontal plane). The video tutorial is an animation with three samples which shows how a cross-section of a 3D structure is created based on a given slicing plane. Example 1 is a Simple structure (Sphere) with a horizontal slicing plane; Example 2 is a Simple structure (Cylinder) with an oblique slicing plane; and Example 3 is an Embedded structure (Cylinder insider a Sphere) with a vertical plane.

We also categorized the questions based on whether they would have seen them in the video tutorial: Level 1 (14 questions) are 3D structures that have a sphere or cylinder in

them. Level 2 (30 questions) are the remaining questions. Additionally, all participants saw the Simple sphere slicing example shown in the tutorial (the sphere example is used as a test question for a validity check).

4.1.2 Experiment 1 Method

Participants: We recruited 40 participants from Mechanical Turk and 75 undergraduate students to take our test (totally 115 participants: 36 females, 77 males; 2 students did not indicate their gender). These participants covered a range of education, 3D Modeling experience, age, and gender.

Procedure: Participants completed the test online. At first, they randomly saw either the written instructions or watched the two minutes video tutorial. They then completed the test in the order of the categories given. Questions are randomized within each set, and participants saw a maximum of 30 questions across the entire question set. At the beginning of each set there is one warm-up question followed by its answer; participants were shown the correct answer or congratulated on getting the correct answer as appropriate. At the end of the test participants provided answers to demographics questions (age, gender, level of education, level of experience working with 3D models and cross-sections, and a self-evaluation of skill). We asked participants to skip the questions rather than provide incorrect information if they did not wish to answer them.

4.1.3 Experiment 1 Results

4.1.3.1 Effects of 3D Stimuli and Animation

To check the effects of "3D stimuli and animation", we computed our participants average score for those questions that are similar to the SBS test questions (30 question of Category 1 from Simple, Joined, and Embedded structures). We then compared our test results with the SBS test results. Overall, the mean proportion of correct items on our test (30 questions) was 0.72 (SD = 0.20), The mean proportion of correct items on the SBS test was M= 0.54 (SD = 0.22) [20]. Our participants significantly outperformed the SBS test participants ($t = 5.342$, $p < 0.0001$). Similar to the SBS test, the

highest mean performance was on Embedded Orthogonal items (M = 0.88, SD = 0.25). However, lowest mean performance was on Joined oblique items (M = 0.56, SD = 0.24), which were significantly more difficult than Embedded orthogonal ones (t = 4.305, p < 0.0001). In both tests, oblique problems were more difficult than orthogonal problems (see Figure 4.4).

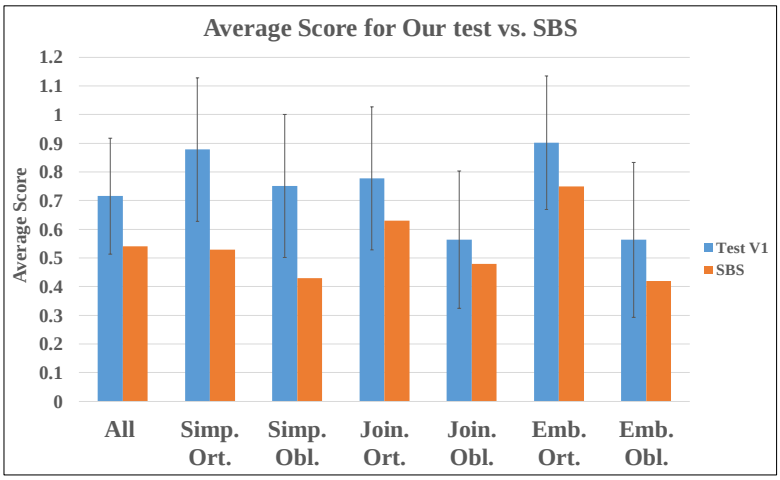


Figure 4.4: Mean proportion of correct items: Our test versus the SBS test.

4.1.3.2 Patterns of Wrong Answers

For the "patterns of wrong answers", we analyzed the average of the four answer choices (correct, egocentric, combination, and alternate) across the 30 test items of Category 1 questions (similar to the SBS test questions). Overall, our participants outperformed SBS test participants. However, we can see a similar pattern of errors (Figure 4.5). The most frequently chosen wrong answer was the egocentric answer. Combination and alternate answers were chosen less frequently.

4.1.3.3 Effects of Shape Complexity and Orientation of Slicing Plane

We conducted a within subjects, repeated measures ANOVA to determine the contribution of "3D object complexity and orientation of slicing plane" to performance on our

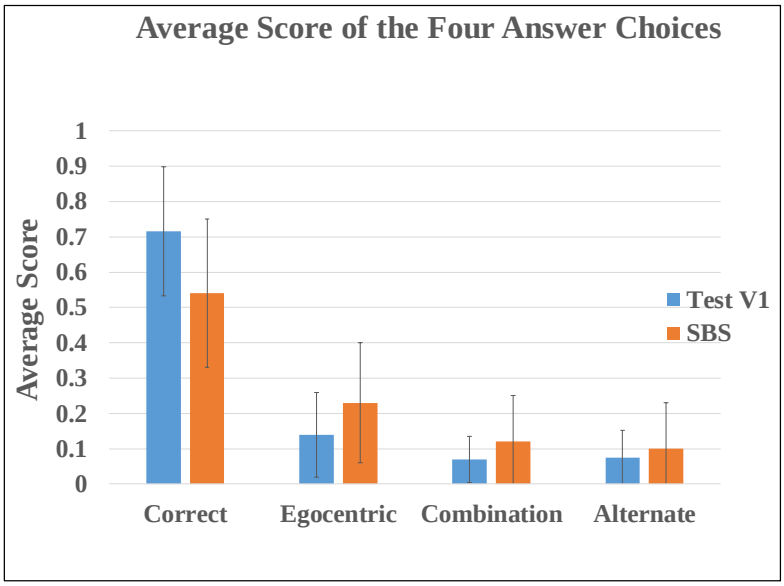


Figure 4.5: Mean proportion of four answer choices on our test versus the SBS test.

test. There was a significant main effect of structure type ($F(3, 111) = 8.755$, $p < 0.0004$) and a significant main effect of slicing plane ($F(1, 114) = 14.753$, $p < 0.0004$). Across the four types of structure, performance was higher on orthogonal than on oblique slicing planes. The lowest performance was on oblique Biological shapes. Within orthogonal items, performance was highest on Embedded, followed by Simple and Joined and then Biological structures. Within oblique items, performance was highest on Simple, followed by Embedded, Joined and finally Biological structures (see Figure 4.6).

4.1.3.4 Same Objects in Different Question Categories

We reused some of the 3D structures of Category 1 in Category 2 and 3 questions. Not all objects appeared in all three categories (e.g., a Simple pyramid was used in Category 2, and a Joined structure of a cone and a cylinder was used in both Category 2 and 3). Figure 4.7 shows the average score for the questions with the same items in each category. The scores varied across the categories and there is no clear pattern. For example, for the Joined structure of two boxes, the score for Category 2 was much lower than Category 1; while for the Joined structure of the cylinder and cone, the score for

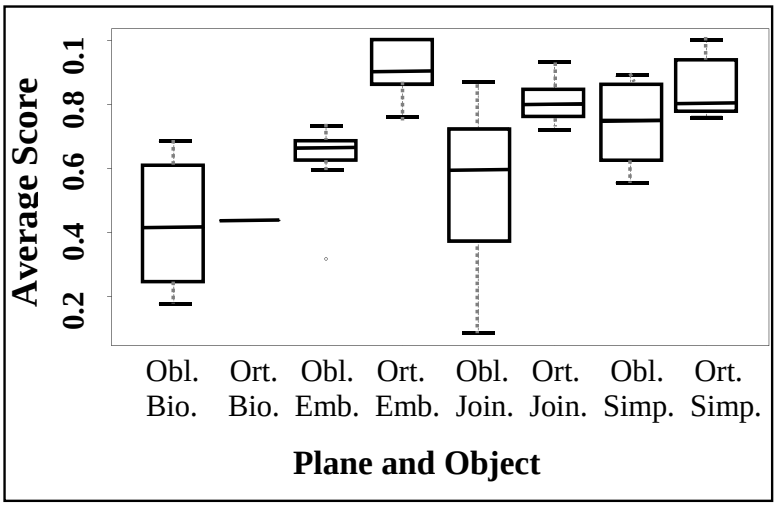


Figure 4.6: Effects of object complexity and plane orientation.

Category 2 was higher than Category 1.

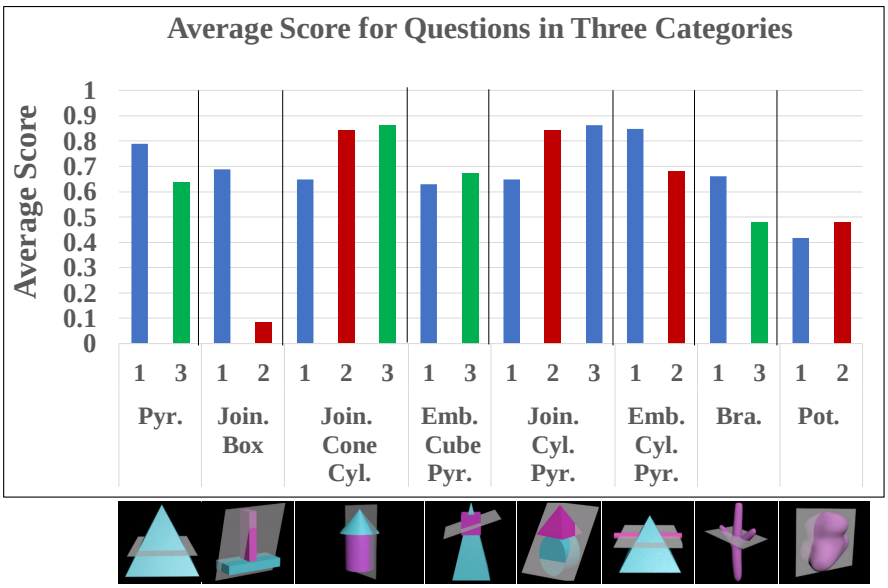


Figure 4.7: Average score for questions with same 3D structures in the three categories.

4.1.3.5 Gender Differences

To check the effects of "gender differences", we separately computed the average score for male and female participants. We then conducted a simple one-way ANOVA to determine the contribution of "Gender" to performance on our test. Overall, the average score for male participants was 0.71 (SD = 0.16). The average score for females was 0.62 (SD = 0.16). Therefore, our male participants significantly outperformed female participants ($F(1,111) = 6.078$, $P = 0.0152$). As shown in Figure 4.8, for both males and females, the highest mean performance was on Simple Orthogonal items: males (M =0.93, SD = 0.15), and females (M = 0.86, SD = 0.21). The second highest performance for both males and females was for Embedded Orthogonal objects: males (M = 0.92, SD = 0.20), and females (M = 0.86, SD = 0.21). Lowest mean performance for males was on Biological Oblique items (M = 0.47, SD = 0.25), but for women lowest mean performance was for Biological Orthogonal items (M = 0.43, SD = 0.25). In summary, males significantly outperformed females on Simple and Embedded items, but not on Joined and Biological ones.

4.1.3.6 Effects of Video Tutorial

To check the "effects of video tutorial", participants were randomly assigned to either watch the video tutorial or read the written instructions (66 participants watched the video tutorial, and 49 participants read the written instructions). The average score of all three question categories for those who watched the video tutorial was slightly higher (M = 0.65, SD = 0.17) than those who just read the written instructions ((M = 0.61, SD = 0.20)). However, this result is not statistically significant. Figure 4.9 shows the average of correct answers split out by ones that were seen in the tutorial (Level 1) and those that were not (Level 2). Again, there is a difference in means, but the results are not statistically significant.

4.1.3.7 Background Questionnaire Analysis

Analyzing the participants' answers to the background questionnaire helped us get some insights about their perceived level of experience in 3D modeling and cross-section familiarity, perceived benefit of the test on cross-section understanding, perceived success

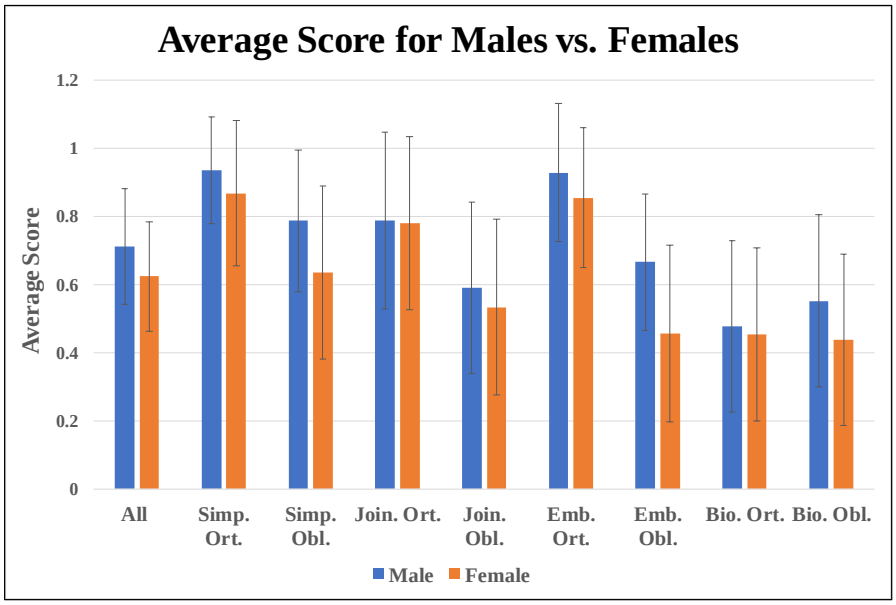


Figure 4.8: Mean performance by gender showing interactions of structure with slicing plane and with gender (n=113, the data of two participants were not included because they did not report their gender).

on completing the test, and perceived difficulty of the test. As shown in Figure 4.10 participants indicated that on average they were 41% familiar with the concept of the cross-sections. They ranked their 3D modeling experience to be 28%. The average perceived benefit of the test on enhancing the cross-section understanding was 64%, and the average difficulty of the test was 51%. Finally, participants claimed that they were 66% successful in providing correct answers to the test.

We are also interested to know if there is any relationship between each of the above attributes and the average score of participants. Figure 4.11 shows correlation of level of experience and familiarity to cross-section with the average score for each participant. Although technically there is a positive correlation between perceived familiarity to cross-sections and average score, the correlation between the variables is weak ($r = 0.2046$, $p = 0.0282$). Similarly, the correlation between 3D modeling experience and average score is weak and not significant ($r = 0.1118$, $p = 0.234$).

Figure 4.12 shows correlation of the average score for each participant with perceived

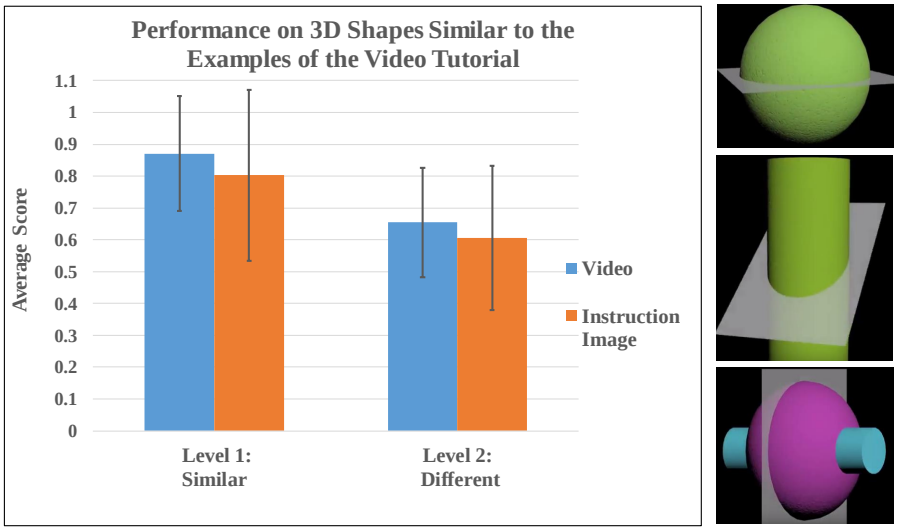


Figure 4.9: Left: Average score (Level 1 and 2) for the the video tutorial versus the written instructions. Right: 3D objects seen in the tutorial.

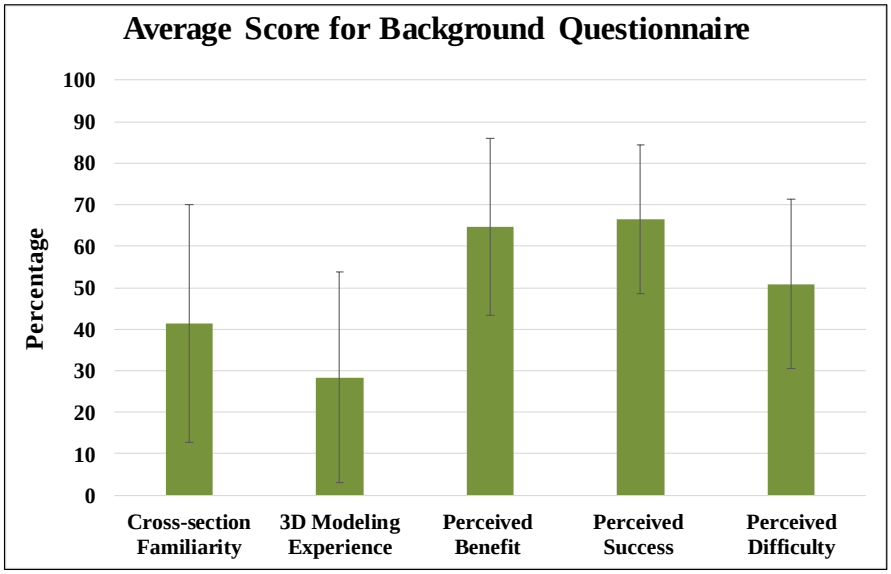


Figure 4.10: Average score for background questions.

benefit (how the test was helpful for participants to understand cross-sections), perceived success (how participants think they were successful in answering the test questions),

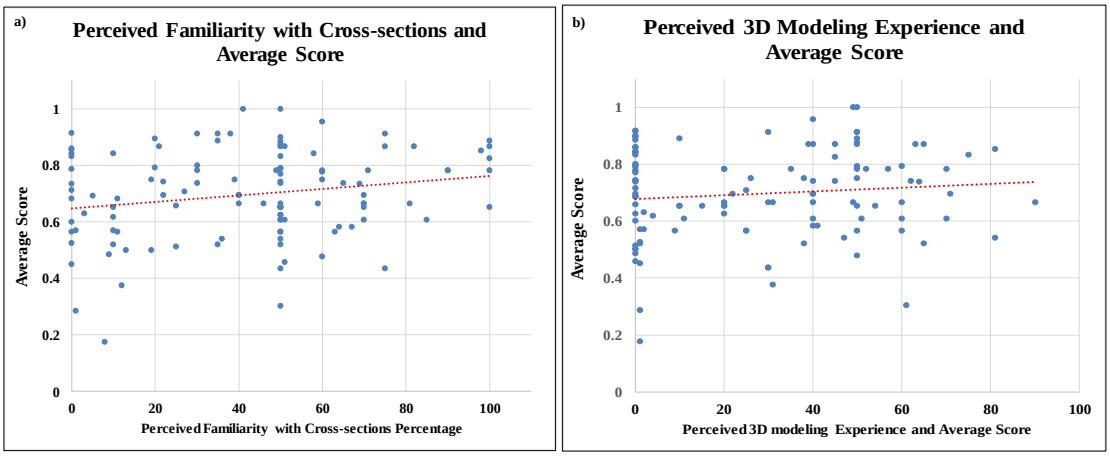


Figure 4.11: Correlation between average score and a) perceived familiarity with cross-section, and b) 3D modeling experience is week.

and perceived difficulty (how difficult the test was for participants). Again, there is no significant correlation between average score and both perceived benefit ($r = -0.058$, $p = 0.5376$), and perceived difficulty ($r = 0.14$, $p = 0.13$). However, there is significant positive correlation between average score and perceived success ($r = 0.38$, $p < 0.0001$).

4.1.4 Discussion for 2D Cross-section Understanding Test 1

[RQ1] In summary, our 3D stimuli improved the participants ability to correctly identify the cross-section, but the overall pattern of correct and incorrect responses remained the same, indicating that part of the difficulty of the task was visualizing the 3D structure of the objects from the 2D image.

[RQ2] Questions with oblique slicing plane were more difficult than items with orthogonal planes. This result is consistent with SBS test results. We further investigate other attributes contribute on difficulty (see Section 4.2).

The animation we added to the 3D structures (rotation around y axis) helped participants to correctly visualize the 3D structures and cross-sections especially for orthogonal slicing planes. It is possible that letting the participants interactively move the camera around would further improve their score, or that simply showing two views (instead

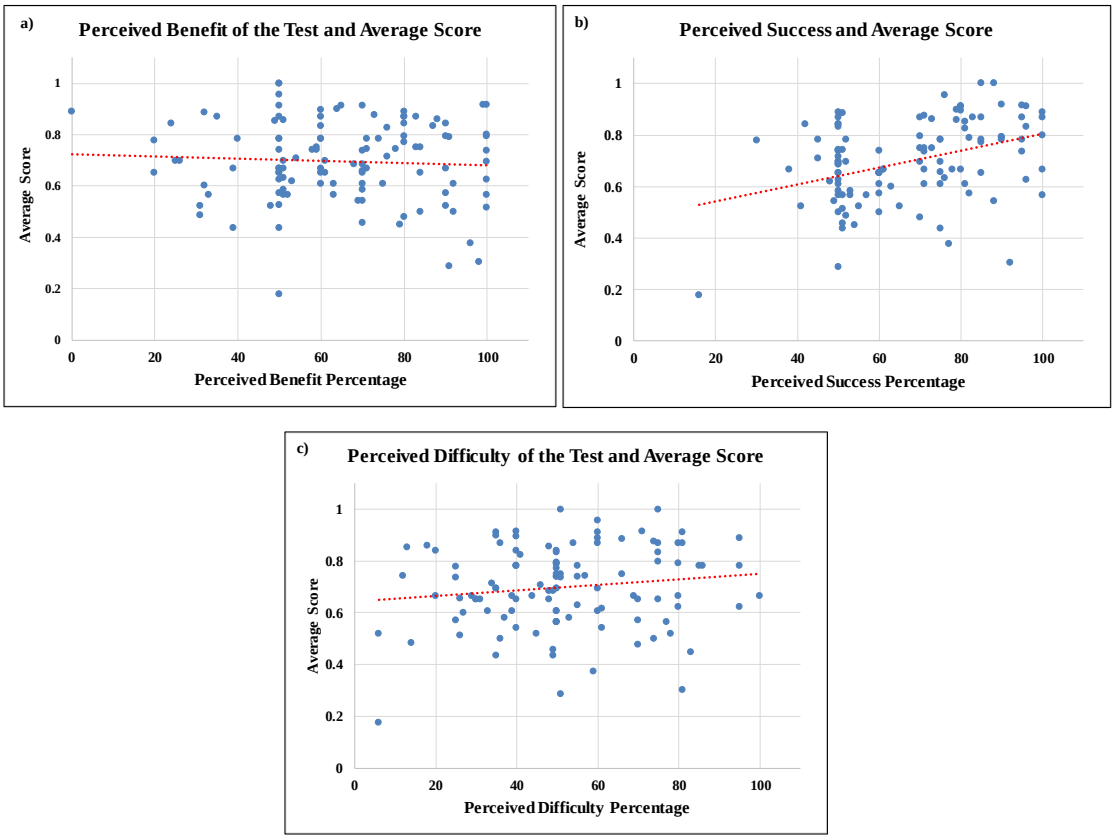


Figure 4.12: Correlation between average score and a) perceived benefit of the test, and c) perceived difficulty is week. b) There is a significant positive correlation between average score and perceived success.

of one) would be just as useful. As with all camera manipulation, there is a trade-off between viewpoint selection and controllability: free viewpoint selection can result in an inexperienced user getting lost or not being able to navigate to their desired viewpoint. Because we wanted to use an on-line survey with the same stimulus for all participants, we compromised and used videos with a fixed rotation direction. In general, the more information sources to evaluate (e.g., number of cross-sections or objects) the better the performance. For example, in Embedded objects (two geometric shapes in specific spatial relationships to each other), a participant could use the additional visual infor-

mation inherent in the answer choice shapes to eliminate the incorrect cross-sections. This strategy is not available for Simple or Biological structures.

[RQ2] For the Biological structures we deliberately chose shapes that were difficult and not just bumpy versions of the existing geometric stimuli (e.g., a tubular shape is essentially a cylinder, and a tapering shape is a cone). Analysis of the questions suggests that Biological structures were more difficult than geometric structures and the challenges were different. Participants had difficulties matching the bumps in the biological structures (e.g., a bumpy shape with a hole in the middle, or a potato structure with similar aspect ratios but different bump patterns), particularly when the camera did not stop when facing the plane. We argue that the viewpoint orientation, and the ability to pause the rotating 3D structure and match up the bump and plane, is a key to provide correct answers to questions with Biological/Organic structures. In addition, similar questions in different categories had different score patterns. This suggests that inverting/modifying the same question (e.g., a question in Category 2 versus Category 1) is not testing the same skill set.

[RQ3] We expected that the video tutorial would be more helpful than the written instructions. However, study results show that static and video tutorials were both equally effective. This may indicate that the tutorial examples were too trivial. There is also a chance that some of the participants did not pay enough attention to the entire two minutes video tutorial .

[RQ4] Male participants significantly outperformed females on both slicing plane orientations of Simple and Embedded test questions, but not on Joined and Biological one. This results is consistent with the SBS test results [21] that shows SBS male participants significantly outperformed females on both Simple and Embedded test items, but not on Joined figures. We are interested to know what could cause the significant gender differences in performance on Simple and Embedded objects but not Joined and Biological ones.

As described in [21], people can use a variety of strategies on spatial tasks, such as imagistic and analytic strategies (including feature/pattern matching, and task decomposition) [84, 39]. These analytic strategies include task decomposition. For Joined

structures, participants might use a task decomposition strategy to separately consider the possible cross-sections of the two joined objects and eliminate answer choices in which one of the objects is not possible. As Biological shapes were more difficult than geometric shapes, participants might need to use a feature/pattern matching strategy to find the correct answer (e.g., match up the bump and plain). However, for Simple and Embedded structures, participants might need to use more imagistic strategies. Consistent with what the authors of [21] speculated, we also argue that males may have a larger set of spatial strategies (such as imagistic strategies) than females, and for that reason they outperformed females on simple and Embedded, but not Joined and Biological structures.

In conclusion, results from both our 2D Cross-section Test Version 1 and the SBS test suggest that question item difficulty varied along two scales: Orientation of the slicing plane and object complexity. Performance was lower for inferring cross-sections of oblique planes. In our test, inferring the cross-section of the organic shapes was a more difficult task and participants needed to rely on small shape changes and cues to choose the correct answer.

Unfortunately, these two scales are not enough to completely set the level of difficulty of a test item. Considering different 3D spatial/visualization skills necessary for inferring a 2D cross-section, our goal is to define a new range of difficulty and predict the level of difficulty for any given cross-section inferring task (e.g., test question items, or training tasks). We ultimately use this range of difficulty to create our modified test instrument (2D Cross-section Understanding Test Version 2) and our training tool tasks.

4.2 A Novel Classification for Defining Range of Difficulty in 2D Cross-section Understanding

Our goal is to identify and quantify sources of difficulty for inferring 2D cross-sections of 3D structures. This will help us define a range of difficulty which is then used to create more balanced (in terms of structure complexity) 2D cross-section test measures and training tasks. To define a correct range of difficulty, we first need to answer this question: What 3D spatial/visualization skills are necessary for inferring a 2D cross-section?

[88] introduced five components that create spatial skills. These are: 1) Spatial Per-

ception; 2) Spatial Visualization; 3) Mental Rotations; 4) Spatial Relations; and 5) Spatial Orientation. [88] also studied the earlier work of [60] and [93], and focused on a classification scheme for spatial skills (see Figure 4.13). They introduced two main categories of 3D spatial skill: 1) Spatial visualization; and 2) Spatial orientation. Spatial visualization involves the ability to imagine and mentally transform spatial information (e.g., mentally moving objects). Spatial orientation involves the ability to imagine oneself or a configuration from different perspectives (mentally move your viewpoint while the object remains fixed in space) [42].

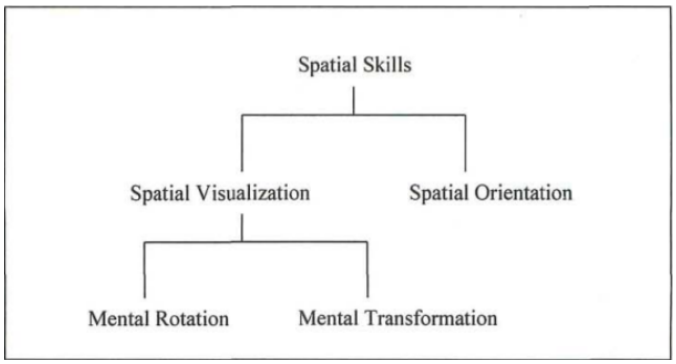


Figure 4.13: Classification of spatial skills [88].

Unfortunately, little is known about the spatial skills, cognitive abilities, and types of information needed to effectively infer 2D cross-sections of 3D objects. [52] introduces an informal process model of the steps involved in a cross-section task as shown in Figure 4.14. The steps include: 1) Observing the external visualization of the object and construct an initial mental representation of the 3D structure. 2) Observing the object from the correct viewpoint (by mentally rotating the object or changing the perspective) 3) Imagining the slicing of the object and removing the part of the object cut by the slicing plane. 4) Inferring what the cross-section will look like (see Figure 4.14).

The above model mainly covers the cognitive steps of a certain cross-section task, inferring the cross-section of a complex 3D object (a tooth), without much elaboration on the required spatial skills to complete the task [51, 50, 41, 52]. To fill this gap and based on the classification of spatial skills (see Figure 4.13), we introduce an extended hierarchy for 2D cross-section understanding to focus on those spatial skills necessary to infer 2D cross-sections of 3D objects (Figure 4.15). The goal of this hierarchy is:

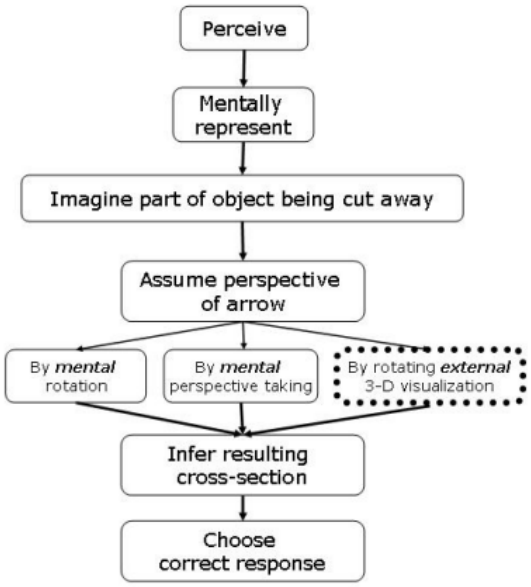


Figure 4.14: An informal model of the processes involved in performing the task of visualizing a cross-section [52].

- To capture 3D spatial skills necessary to infer 2D cross-sections of 3D objects.
- To break down the main task of 2D cross-section understanding into sub-tasks and map each sub-task to the corresponding spatial skill. This will help us understand what spatial skills are required for a certain sub-task.
- Assign each of the questions of the 2D cross-section test instrument version 1 to a leaf sub-task. In this way we can understand what spatial skills and sub-tasks are required to correctly answer a test item.

The key elements of this spatial skill hierarchy include:

- *Spatial Orientation* is a skill needed to assume the correct perspective of the slicing plane with respect to the view/location of the slicing plane and the 3D object (Orthogonal/45 degree/arbitrary; Center aligned/non-center aligned).
- *Mental Rotation* is a skill needed to identify the correct perspective of a cross-section by mentally rotating an object/viewpoint.

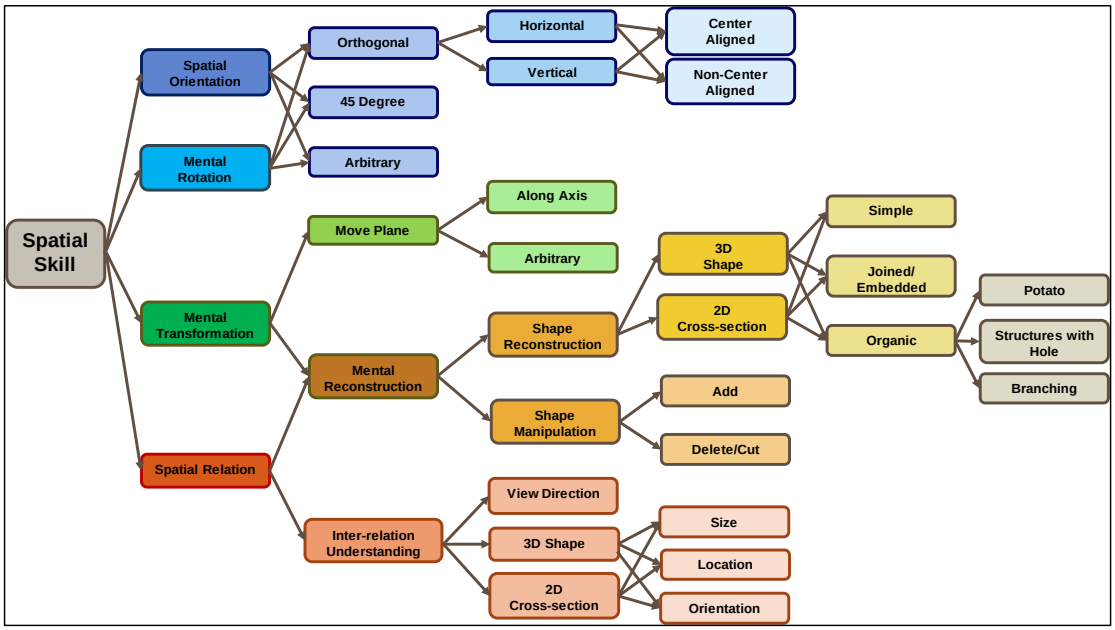


Figure 4.15: Extended hierarchy of the spatial skills involved in performing the task of inferring a 2D cross-section.

- *Mental Transformation* is a skill needed to mentally move a plane or to reconstruct the object (e.g., to correctly imaging part of the object that has been cut away by a given slicing plane). Mental reconstruction depends on the object shape (2D/3D).
- *Spatial Relation* is a skill to mentally compare different cross-sections and their relationship to the 3D structure.

We used the extended hierarchy to categorize the questions of test version 1. In addition, we considered participants' average score for each question item as an estimate for question difficulty. We defined seven basic difficulty tags and assigned those tags to each of the questions: 1) very easy (average score is above 90%); 2) easy (average score is above 80%); 3) easy-medium (average score is above 70%); 4) medium (average score is above 60%); 5) medium-hard (average score is above 50%); 6) hard (average scoreis above 30%); and 7) very hard (average score is below 30%). We then mapped each of the test questions to one of the leaves of the hierarchy (see Figure 4.16).

While this approach helped us categorize the test questions and understand what

spatial skills were required to correctly answer them, it is still challenging to assign a certain level of difficulty to each question item of the test. Using the hierarchy, it is impossible to justify why a certain question is tagged to be easy or difficult; and to highlight all the sources of difficulty of a certain question item. To overcome this problem, by using the elements of the hierarchy, we introduce a novel classification for defining a range of difficulty in 2D cross-section understanding.

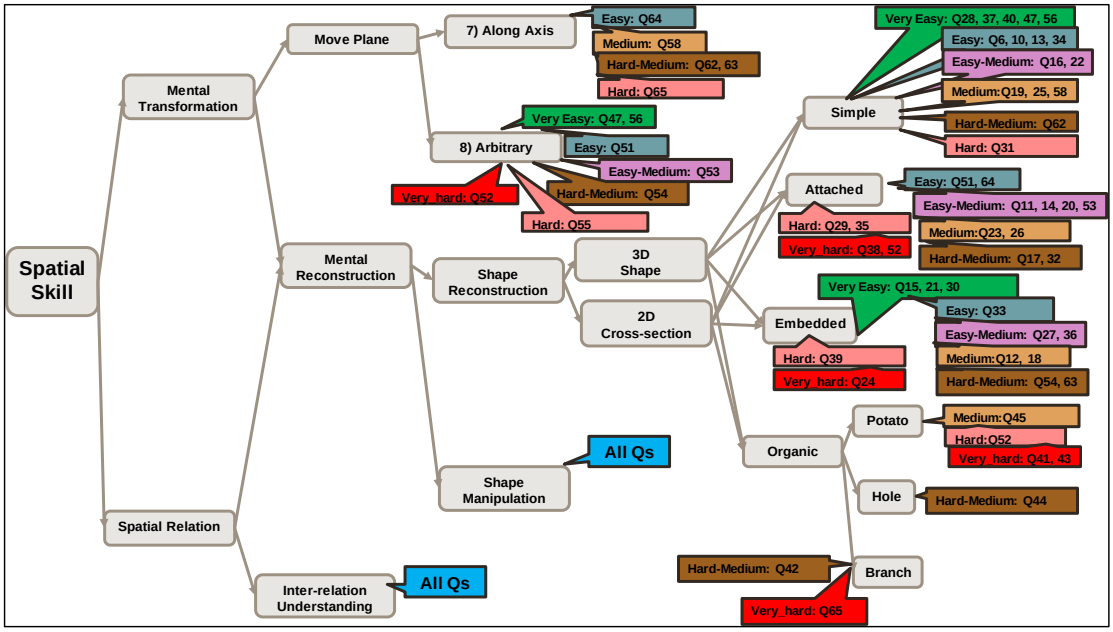


Figure 4.16: 2D cross-section understanding hierarchy example along with questions and difficulty tags for “Mental Transformation” and “Spatial elation” spatial skills.

4.2.1 Range of Difficulty for Inferring 2D Cross-sections

Going back to our 2D Cross-section Understanding Test Version 1, we focus on the question with the lowest average score. What made this question the most difficult one for our participants? Let’s consider its characteristics: The question consists of an organic shape with an oblique plane. Also, the viewpoint with respect to the plane/3D shape is not orthogonal (See Figure 4.17). We propose that in addition to the orientation of the object and the slicing plane, it is important to consider viewpoint difficulty with

regards to both the plane and the 3D object.

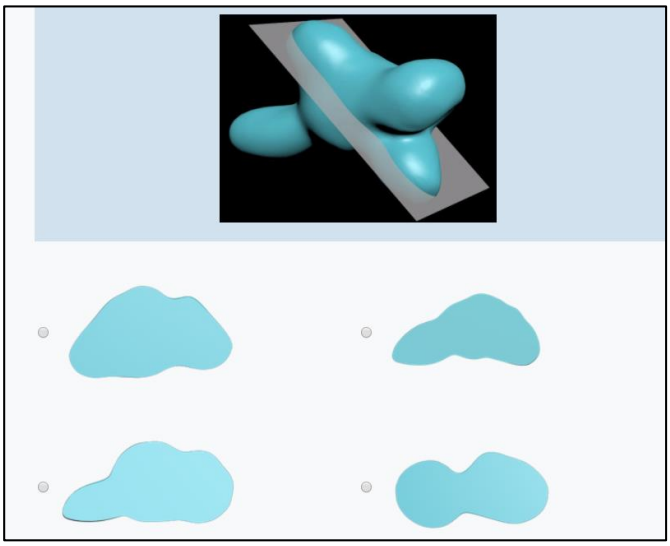


Figure 4.17: Most difficult question of Test Instrument Version 1 (based on average score). Viewpoint with respect to both the plane and the 3D object is not orthogonal. Also, the 3D structure itself and 2D representation form cut-away are complex objects.

Using our extended hierarchy (see previous section), we built a range of difficulty for inferring 2D cross-sections of 3D objects based on four spatial attributes: 1) Viewpoints (based on spatial orientation skill of the extended hierarchy); 2) Mental rotation/transition (based on mental rotation and mental transformation); 3) 2D object representation from cut-away (based on mental transformation and spatial relation); and 4) 3D object representation (based on mental transformation and spatial relation).

In this section, we describe each of these attributes and demonstrate how they influence difficulty. Then, we define difficulty levels for each of the four attributes, starting from Level 0 or "Base" difficulty. The level of difficulty increases based on certain aspects compared to the "Base" level. Table 4.2 summarizes all the attributes and levels of difficulty along with examples. We now explain each of these attributes in detail.

Table 4.2: Summary of our novel approach to define a range of difficulty: 2D cross-section understating attributes and levels of difficulty.

Attribute		Base:0	Range of Difficulty
Viewpoint	**View Point -Fixed**	**Base 0: All 3 attributes are true (3/3)** 1. Simple object AND 2. Viewpoint wrt. the plane is orth. AND 3. Viewpoint wrt. the object is orth. to the major/minor axis	**Level 1: Only one of the base attributes is false (2/3)** Example: Complex object; AND Viewpoint wrt the plane is Orth; AND Viewpoint wrt the object is orth. to the major/minor axis; **Level 2: Only one of the base attributes is true (1/3)** Example: Simple object; AND Viewpoint wrt the plane is NOT Orth; AND Viewpoint wrt the object is NOT orth. to the major/minor axis. **Level 3: None of the base attributes is true (0/3)** Example: Complex object; AND Viewpoint wrt the plane is orth. AND - Viewpoint wrt the object is orth. to the major/minor axis.
	View Point -Rotating	**Base 0: All 2 attributes are true (2/2)** 1. Simple object AND 2. Simple Rotation* (Having viewpoint rotation around major/ minor axis of the object OR rotation around one of the axes of the plane)	**Level 1: Only one of the base attributes is true (1/2)** Example: Complex object; AND Simple Rotation **Level 2: None of the base attributes is true (0/2)** Complex object AND Arbitrary rotation;
	View Point -Freeform	**Base 0**: Get to the Base viewpoint fixed AKA **"Simplest Viewpoint"** with **0 mental rotation** (See View Point-Fixed Base:0 for the attributes)	**Level 1:** Get to the Simplest Viewpoint with only 1 mental rotation **Level 2:** Get to the Simplest Viewpoint with 2 mental rotations **Level 3-5:** There does not exist a Simplest Viewpoint (e.g., Object is complex): In this case complexity increases if we have no bases rotations: **Level 3:** Get to a *Simpler Viewpoint with 0 mental rotation **Level 4:** Get to a Simpler Viewpoint with 1 mental rotation **Level 5:** Get to a Simpler Viewpoint with 2 mental rotations *Simpler Viewpoint: When we have a complex object but Viewpoint wrt. the plane is orth. AND Viewpoint wrt. the object is orth. to the major/minor axis
Mental Rotation/Translation (Adjust the plane to complete a task)		**Base 0:** Complete the assigned task with 0 translation AND 0 rotation of the plane wrt to the object	Level of complexity increases with increasing the number of translation and rotations
2D Object Representation from Cut-away		**Base 0: All Attributes are true (2/2)** 1. From Primitive 3D object 2. From an orthogonal plane wrt object (parallel to the cross-section) Example: Circle from a sphere with horizonal plane	**Level 1: Only one of the base attributes is true (1/2)** Example: oval from a cylinder with non-orthogonal plane **Level 2: None of the base attributes is true (0/2)** Example: Branching structure with non-orthogonal plane
3D Object Representation		**Base 0:** Primitive Objects OR Symmetric simple organic (e.g. symmetric potato) shapes	**Level 1:** Attached/Nested Primitive objects **Level 2:** Symmetric nested organic objects (like potato with hole) **Level 3:** Asymmetric simple organic objects **Level 4:** Asymmetric nested organic objects

4.2.1.1 Viewpoint Difficulty

We define that the level of difficulty for understanding a 2D cross-section depends on the viewpoint from which a 3D object is observed. In 2D cross-section understanding, we are dealing with two objects: a 3D structure and a slicing plane. Therefore, we consider viewpoint with regards to both objects. In addition, the shape complexity of a 3D structure influences the viewpoint difficulty. We have three different types of viewpoint:

1. **Fixed Viewpoint**: When there are only static viewpoints and there is no inter-

action with the objects (e.g., static image of a 3D structure and slicing plane).

2. **Rotating Viewpoint:** When objects are rotating (e.g., animated structures rotating along the y axis similar to our test questions). There is no interaction with the objects (e.g. controlling the direction and speed of the rotation).

3. **Freeform Viewpoint:** When we can interact with objects and change the viewpoints freely (e.g. change from top viewpoint to perspective viewpoint).

We now describe each of these viewpoints, starting with "Fixed Viewpoint".

Fixed Viewpoint: To create the level of difficulties, we start with Level 0 or "Base". To be in level 0 for "Fixed Viewpoint", three attributes should be met:
1) The 3D structure is a simple object. By simple object we mean the 3D object is a primitive object, it is an extrusion of a cross-section, and it is symmetric; 2) Viewpoint with respect to the plane is orthogonal; and 3) Viewpoint with respect to the object is orthogonal to the major/minor axis.

If all the above attributes are true, then we are in "Base Fixed Viewpoint" or the simplest possible viewpoint. We then created three other levels of difficulty for "Fixed Viewpoint". The higher the level of difficulty goes the more complex the task of cross-section understanding becomes.
Level 1: Only one of the base attributes is false. For example, when the object is not simple; but viewpoint with respect to both the object and plane is orthogonal.
Level 2: Only one of the base attributes is true: For example, when there is a simple object, but viewpoint with respect to the plane and object is not orthogonal.
Level 3: None of the base attributes are true. This is the case when the object is not simple; and viewpoint with respect to both the plane and object is not orthogonal.

For our 2D cross-section understanding test instruments we consider the "Fixed Viewpoint" to be the one when the 3D objects stop rotating for a few seconds. Figure 4.18 shows the examples of Level 0 (Base) to Level 3 of difficulty.

Rotating Viewpoint: There are two attributes that define Level 0, or "Base", for "Rotating Viewpoint":
1) The 3D structure is a simple object.

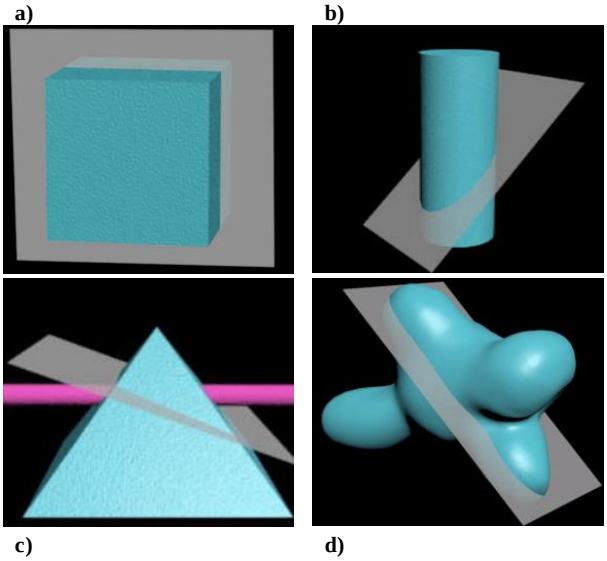


Figure 4.18: Four examples to cover "Fixed Viewpoint" difficulty: a) Level 0 or "Base": The object is simple (cube); Viewpoint with respect to both the plane and the object is orthogonal. b) Level 1: The object is simple (cylinder); Viewpoint with respect to the object is orthogonal, but viewpoint with respect to the plane is not orthogonal. c) Level 2: The object is not simple (Embedded); Viewpoint with respect to the object is orthogonal, but viewpoint with respect to the plane is not orthogonal. d) Level 3: The object is not simple (potato); Viewpoint with respect to both the plane and the object is not orthogonal.

2) The rotation is simple. By "Simple Rotation" we mean: Having viewpoint rotation around major/ minor axis of the object or rotation around one of the axes of the plane. The questions in our test instruments have simple rotation (around y axis).

We then create two other levels of difficulty for rotating viewpoint.

Level 1: Only one of the base attributes is true: For example, the 3D object is not simple, but we have a simple rotation.

Level 2: None of the base attributes are true: when the 3D object is not simple, and we do not a have simple rotation.

All our test instrument questions have a simple rotation. Therefore, the level of difficulty for the "Rotating Viewpoint" is 0 or 1.

Freeform Viewpoint: In "Freeform Viewpoint", we can change the viewpoint freely (e.g., change from top view to left view either by clicking a button or by mouse dragging).

When we have a simple object, we can choose different viewpoints. Only one of those viewpoints is the simplest viewpoint (equal to the "Base" for "Fixed Viewpoint"). To get from any arbitrary viewpoint to the "Base", we need to mentally rotate the simple object/viewpoint. The number of mental rotations needed to go from any viewpoint to the "Base" viewpoint determines the level of difficulty. Therefore, we define Level 0 for "Freeform Viewpoint" as:
Level 0 (Base): Get to the Base Fixed Viewpoint also known as "Simplest Viewpoint" with 0 mental rotation (see Fixed Viewpoint Level 0 for the required attributes).

When we do not have a simple object, then there does not exist a "Base Viewpoint". But again, the level of difficulty changes depending on the number of mental rotations needed to get to the simpler viewpoints. Simpler viewpoints are orthogonal with respect to both the plane and the 3D object.
The levels of difficulty for "Freeform Viewpoint" is:
Level 0 or Base [Simple Object]: Get to the "Viewpoint Fixed Base" also known as "Simplest Viewpoint" with 0 mental rotation.
Level 1 [Simple Object]: Get to the Simplest Viewpoint with only 1 mental rotation.
Level 2 [Simple Object]: Get to the Simplest Viewpoint with 2 mental rotations.
Level 3: Get to a Simpler Viewpoint with 0 mental rotation.
Level 4: Get to a Simpler Viewpoint with 1 mental rotation.
Level 5: Get to a Simpler Viewpoint with 2 mental rotations.

Our test instrument questions do not have "Freeform Viewpoint" but for our training tool, we have provided features to let users change the viewpoints if needed (See Chapter 5).

4.2.1.2 Mental Rotation/Transition Difficulty

One of the other factors that has impact on understanding 2D cross-sections is the mental rotation/transition difficulty. To correctly infer a 2D cross-section, one may need to mentally rotate/move the plane. Mental rotation/transition difficulty depends on how the cross-section inferring task is defined. For our test instrument questions, participants do not need to mentally move the plane. In our training tool design, we will see more

examples of tasks which require plane rotation/transition (see chapter 5). Table 4.3 summarizes some of the tasks that need plane rotation/transition.

"Base" or Level 0 of difficulty is when one can complete the 2D inferring cross-section task with no translation or rotation of the plane. Level of difficulty goes up when the number of translation and rotations increases.

Table 4.3: Tasks that need plane rotation/transition

Tasks	Example
Identify/find cross-section based on a certain part/characteristic of the 3D object	Adjust the plane to capture: Cross-section of the skinniest part of the object/ roughest part of the object.
Locally adjust the plane to create a certain change in the 2D target cross-section	Adjust the plane to change: Cross-section from a circle to an oval in a cylinder.
Place the plane to the cut the 3D structure	Adjust the plane to cut the shape in half [both] vertically and horizontally.

4.2.1.3 2D Object Representation Difficulty

The cross-section of a slicing plane and a 3D structure is a 2D object. For example, cross-section of a plane (with any orientation) with a sphere is a circle. To identify the correct cross-section of a 3D structure, one should understand and accurately visualize the shape of that 2D object. This understanding involves mentally cutting the 3D structure from the location of the plane and imagining the 2D object representation. 2D representation of a primitive object like a cube, is different from a 2D representation of an organic shape like a branching structure. In addition, 2D representation of a cross-section of a cylinder with orthogonal plane (circle) is different from the cylinder with oblique plane (oval). Therefore, the difficulty of mentally understanding the 2D object representation depends on both the 3D structure, and the orientation of the slicing plane.

Level 0 (Base): The difficulty level of 2D object representation is 0 or "Base" if the two below attributes are true:
1) 2D object representation is created from a simple primitive object.
2) 2D object representation is created from an orthogonal plane with respect to the object (parallel to the cross-section).
Circle from a sphere with horizontal plane is an example of Level 0 or "Base" difficulty. We then define two other levels of difficulty:
Level 1: Only one of the base attributes is true: For example, the 3D object is simple, but the plane orientation is non-orthogonal (oval representation from a cylinder with

oblique plane).
Level 2: None of the base attributes are true: when the 3D object is not simple, and the plane orientation is non-orthogonal (2D representation from a branching structure with non-orthogonal plane).

4.2.1.4 3D Object Representation Difficulty

3D object representation also has a great impact on the level of difficulty of inferring cross-sections. It is more difficult to understand and visualize the cross-section of a complex 3D object. For example, visualizing a cross-section of a sphere with an orthogonal plane is more straightforward than the cross-section of a branching shape with an orthogonal plane that cuts the branch point.

3D object representation difficulty is in level 0 or "Base" if the 3D object is a primitive simple (e.g., cube) or a symmetric simple organic (e.g. symmetric potato) shapes. Other levels of difficulty are:
Level 1: When we have attached/nested primitive objects.
Level 2: When we have symmetric nested organic objects (like symmetric potato with hole).
Level 3: When we have asymmetric simple organic object.
Level 4: When we have asymmetric nested organic objects.

4.2.2 Categorizing Questions in Cross-section Understanding Test Version 1 Based on Level of Difficulty

By summing up all the difficulty values for each attribute, we measured difficulty score for each question item in 2D cross-section Understanding Test Version 1. For example, suppose we have: Viewpoint Fixed (VFix)=0, Viewpoint Rotation(VR)=0, Viewpoint freeform (VFree) =0, Mental rotation(MR) =1, 2D Object representation(2D) =0, and 3D object representation(3D) =0. Then difficulty grade for this item is 1. For Test Version 1 we identified 14 Levels of difficulty.
Example: Figure 4.18 shows four question items in four levels of difficulty:

- a) Difficulty score=1, Difficulty level=1: (Vfix=0, VR=0, VFree=0, MR=1, 2D=1,

and 3D=0);

- b) Difficulty score=4, Difficulty level=3: (Vfix=1, VR=0, VFree=1, MR =1, 2D=1, and 3D=0);
- c) Difficulty score=10, Difficulty level=7 (Vfix=2, VR=0, VFree=3, MR =2, 2D=2, and 3D=1);
- d) Difficulty score=15, Difficulty level=10 (Vfix=3, VR=0, Viewpoint freeform =5, MR =2, 2D=2, and 3D=3).

We predicted that performance would be better on questions with lower level of difficulty. To check this, we computed the average score of all participants for each of the questions (from survey results). For example, we had seven questions in Level 1 difficulty. On average more than 84% of participants provided correct answers for all questions of this level. Figure 4.19 shows questions average score and level of difficulty of each question item. As predicted there is a negative correlation, which means that higher average scores go with lower difficulty levels (r = -0.70 , p<0.00001).

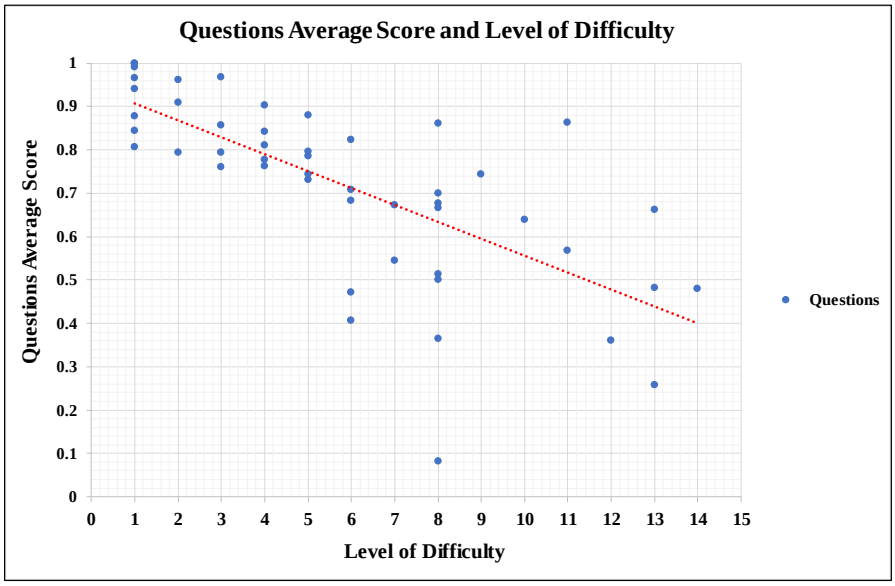


Figure 4.19: 2D Cross-section Test Version 1: Questions Average Score and Level of Difficulty. Higher average scores go with lower difficulty levels (r = -0.70 , p<0.00001).

We use this novel range of difficulty to create a new test instrument with only organic shapes and a fewer number of question items.

4.3 2D Cross-section Understanding Test Version 2

In the previous section we introduced a new approach for categorizing test questions based on our novel range of difficulty. We use this approach to redesign our test items to have questions that: 1) Captures skill sets based on range of difficulty; and 2) Focuses on organic 3D structures instead of primitive ones.

In design, implementation, and evaluation of this test instrument, we investigate the following research questions:

- **RQ1)** What is the relationship between range of difficulty and participants' response to the test questions?
- **RQ2)** What are the sources of difficulty for the participants in inferring 2D cross-sections of different 3D structures?
- **RQ3)** Is a video tutorial on 2D cross-sections more helpful than static/written instructions?
- **RQ4)** Do gender differences effect performance on understanding 2D cross-sections?
- **RQ5)** What are participants' opinions about the test (in terms of perceived success, perceived difficulty, and perceived benefit of the test)?

4.3.1 Test Design and Implementation

We first started by creating the organic 3D models based on 2D and 3D representation difficulty. The gaol is to balance topological/shape complexity of test questions using our range of difficulty. Having Test Version 1 geometric shapes in mind, in an iterative process, we sketched different organic shapes, then used clay to make a real model of the objects, and ultimately chose a few of them for digital implementation. The final 3D objects included symmetric simple structures (potato shapes), asymmetric simple structures (branching shapes), and nested objects (structures with whole). We used 3D

Studio Max [6], ZBrush [71], and JustDrawIt [37] applications to create the digital 3D models. Figure 4.20 shows an example of the process of creating 3D models.

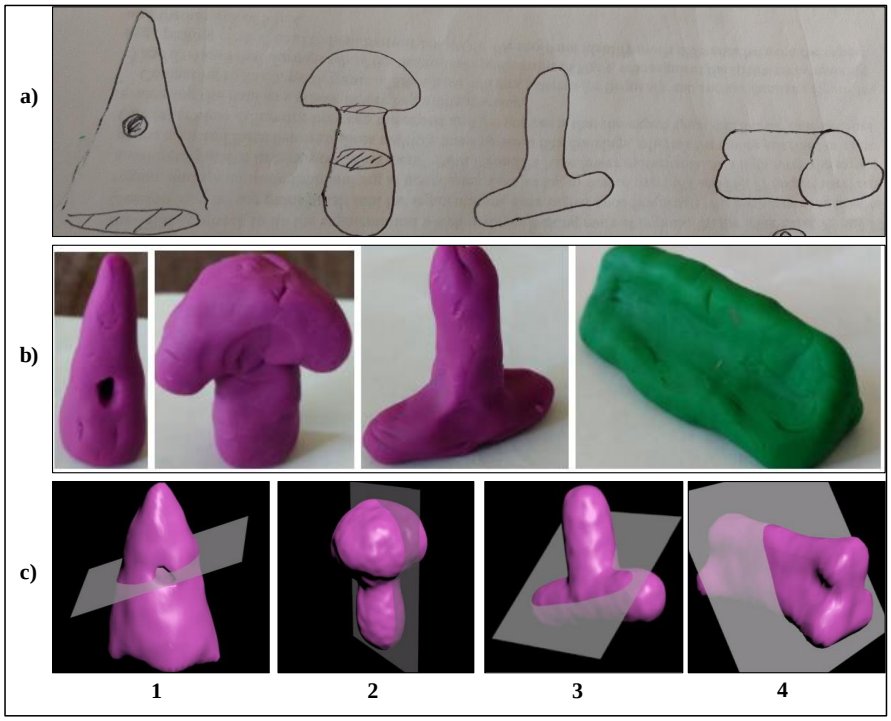


Figure 4.20: a) low level; b) clay; and c) digital implementation of 3D models.

We then created the question items based on the other difficulty attributes: Viewpoint and mental rotation/transition. The 3D structure rotates (around y axis). Since all of the 3D objects are organic structures (not simple), the Viewpoint Rotating difficulty for all the items is 1. For each test item, viewpoint with respect to the objects (3D model and slicing plane) is either orthogonal or non-orthogonal. Similar to 2D Cross-section Understanding Test Version 1, we have three categories of questions:

- **Category 1:** Given a 3D structure and slicing plane, identify the correct 2D cross-section contour (Figure 4.21 a).
- **Category 2:** Given a 2D cross-section for a 3D structure, identify the slicing plane that generated the contour (Figure 4.21 b).

- **Category 3:** Given a 3D structure and multiple slicing planes, identify the valid contour sequence that corresponds to the slices (Figure 4.21 c).

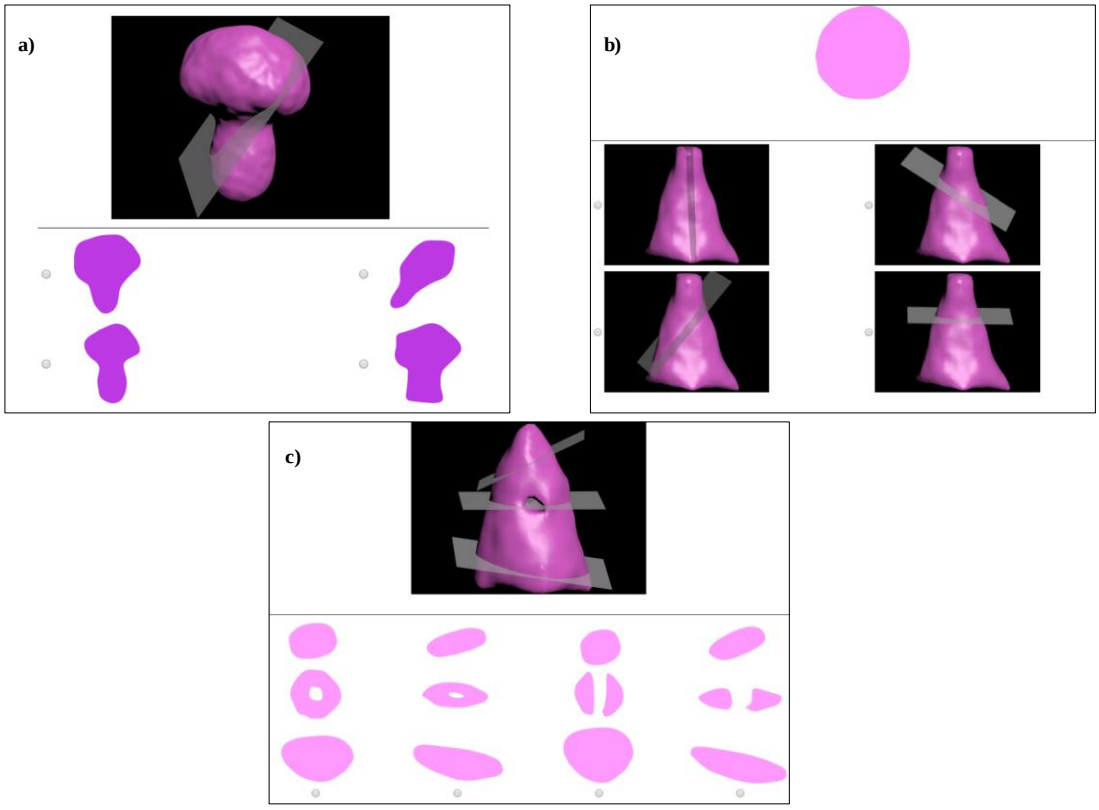


Figure 4.21: Sample questions from 2D Cross-section Understanding Test Version 2. a) Category 1; b) Category 2; and c) Category 3.

By combining the shape complexity of 3D models with the viewpoint orientation and mental rotation/transition, we created 22 questions items in 10 levels of difficulty. 12 questions are of Category 1, six questions are of category 2, and four questions are of category 3. We implemented the test questions in Qualtrics. Figure 4.22 shows question items in three different levels of difficulty.

Except for questions of type 2, each item has three types of wrong answers: 1) "Egocentric" wrong answer (Figure 4.23 a) represents a shape that participants might visualize if they failed to change their view perspective relative to the slicing plane of the criterion

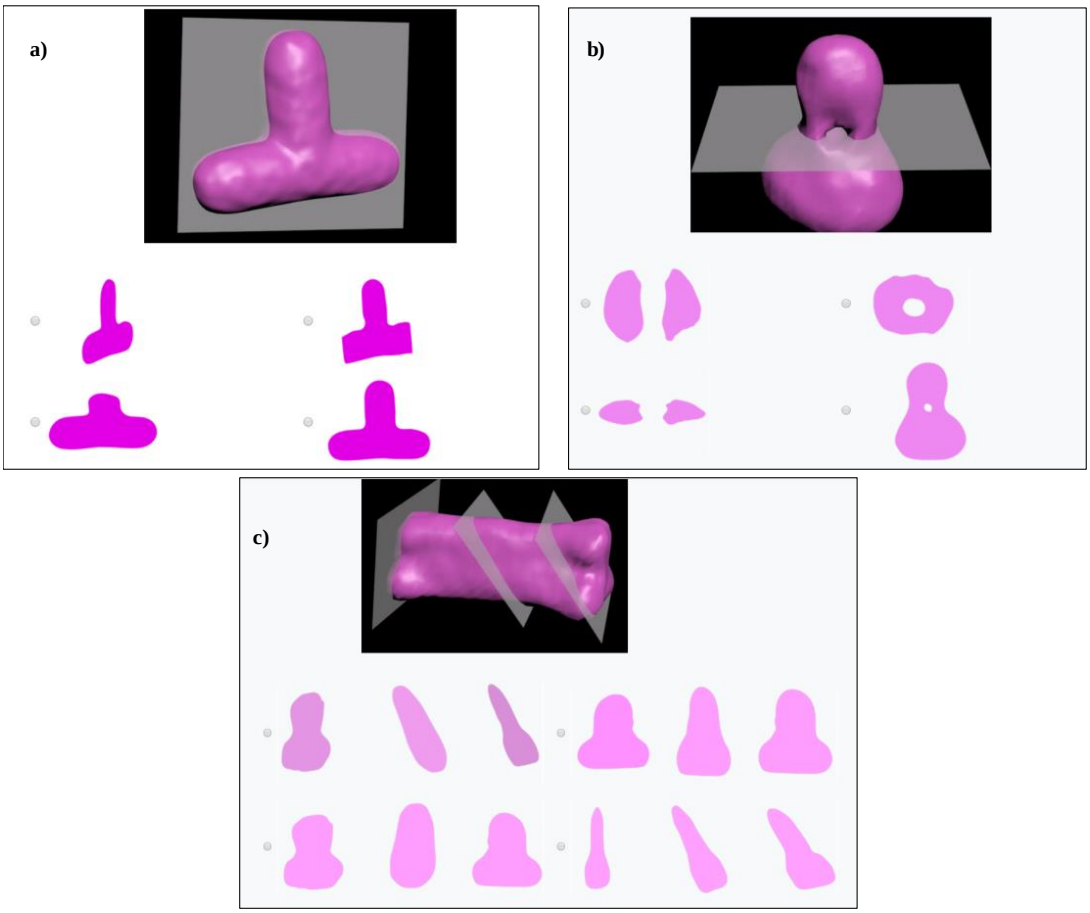


Figure 4.22: Sample questions with different levels of difficulty. a) Level 1 of difficulty (simple organic shape; viewpoint orthogonal with respect to both the plane and object; vertical plane). b) Level 5 of difficulty (organic shape with hole; viewpoint not orthogonal to the object; horizontal plane). c) Level 10 of difficulty (asymmetric organic shape; planes with different orientation; viewpoint not orthogonal with respect to both object and plane; multiple mental transitions/rotations needed to accomplish the task.

figure. 2) “Not possible” wrong answer (Figure 4.23 b) shows an impossible cross-section (independent of the orientation of the slicing plane, it is not possible to have such cross-section). 3) “Alternate” wrong answer (Figure 4.23 c) shows another possible slice of the test figure; 4) Figure 4.23 d shows the correct answer.

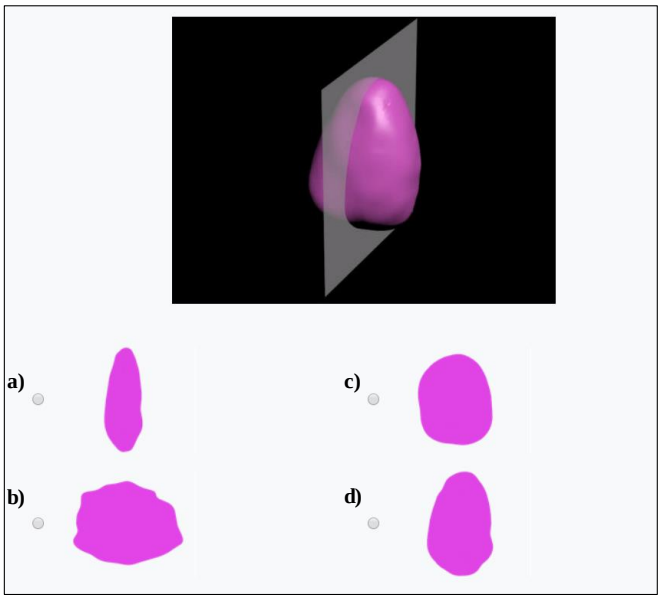


Figure 4.23: A Simple potato shape test item and the four answer choices: a) Egocentric; b) Not possible; c) Alternative; and d) Correct answer.

4.3.2 Experiment 2 Method

We conducted user studies to validate our 2D cross-section test instrument version 2.

Participants: We recruited 30 participants from Mechanical Turk (15 males, and 10 females; 5 participants did not indicate their gender). These participants covered a range of education, 3D Modeling experience, age, and gender.

Procedure: Same as test version 1, participants completed the test online. At first, they randomly saw either the written instructions or watched the two minutes video tutorial. They then completed the test in the order of the categories given. Questions are randomized within each set. At the beginning of each set, there is one warm-up question followed by its answer; participants were shown the correct answer or congratulated on getting the correct answer as appropriate. At the end of the test participants provided answers to optional demographic/background questions (age, gender, level of education, level of experience working with 3D models and cross-sections, and a self-evaluation of

skill/success).

4.3.3 Experiment 2 Results

4.3.3.1 Analysis based on Ranges of Difficulty

We predicted that performance would be higher on questions with lower level of difficulty. To check this, we computed the average score of all participants for each of the questions. As predicted there is a strong negative correlation, which means that the higher average scores go with lower difficulty levels ($r = -0.84$, $p < 0.00001$). Figure 4.24 shows the relationship between questions average score and level of difficulty.

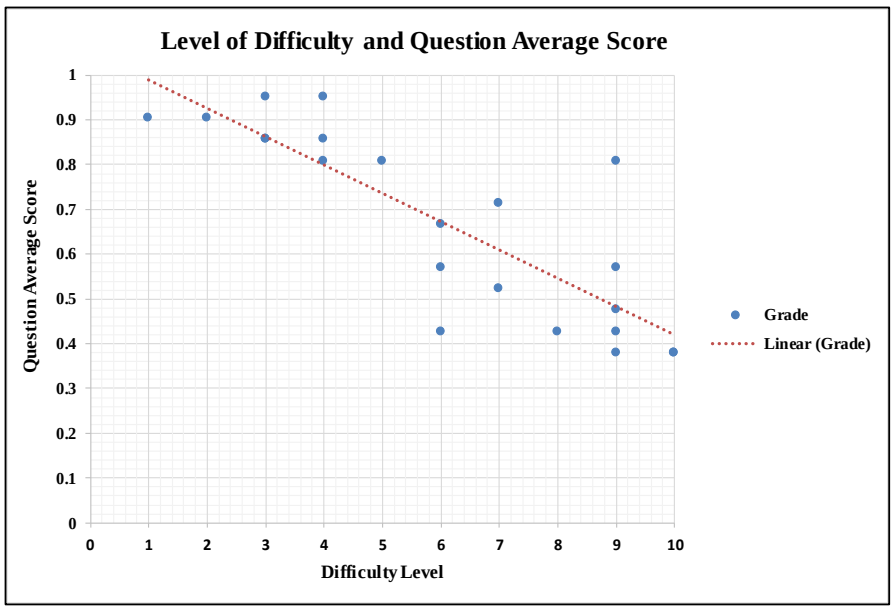


Figure 4.24: Relationship between level of difficulty and question average score. We have 10 levels of difficulty. As the level of difficulty increase, the average score decreases.

We conducted a within subjects, repeated measures ANOVA to determine the contribution of "Question category", "Orientation of slicing plane", "Viewpoint Difficulty Level" and "3D object representation" to performance on our test. There was a significant main effect of slicing plane orientation ($F(1, 28) = 13.29$, $p < 0.0006$) and a very significant main effect of viewpoint difficulty level ($F(2, 27) = 47.08$, $p < 0.00001$).

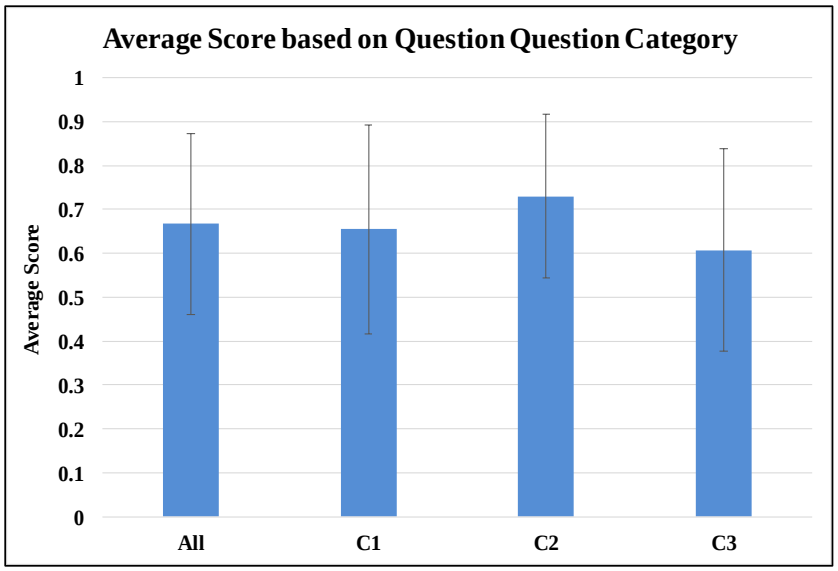


Figure 4.25: Average score based on question category. On average (but not significantly) Category 2 has the highest performance, and Category 3 has the lowest.

However, we cannot conclude there is any significant effect of question category and 3D object representation ($p > 0.6$).

In addition, we separately checked the effect of the plane orientation and question category on average scores. As Figure 4.25 shows, the highest performance was for Category 2 ($M = 0.73$, $SD = 0.18$), while the lowest mean performance was for Category 3 questions ($M = 0.61$, $SD = 0.23$). If we focus on plane types (see Figure 4.26), we can see that the questions with oblique plane were significantly more difficult than questions with orthogonal (specifically vertical) planes ($t = 5.5$, $p<0.001$). The results are consistent with our prediction and the previous test results (both SBS and Test Version 1).

Combining plane orientation with question category, we computed average score for each of the participants. As Figure 4.27 shows, across the three categories, oblique questions are more difficult than orthogonal questions. Questions of Category 1 with oblique plane have the lowest performance, while Questions of Category 2 with vertical plane have the highest performance.

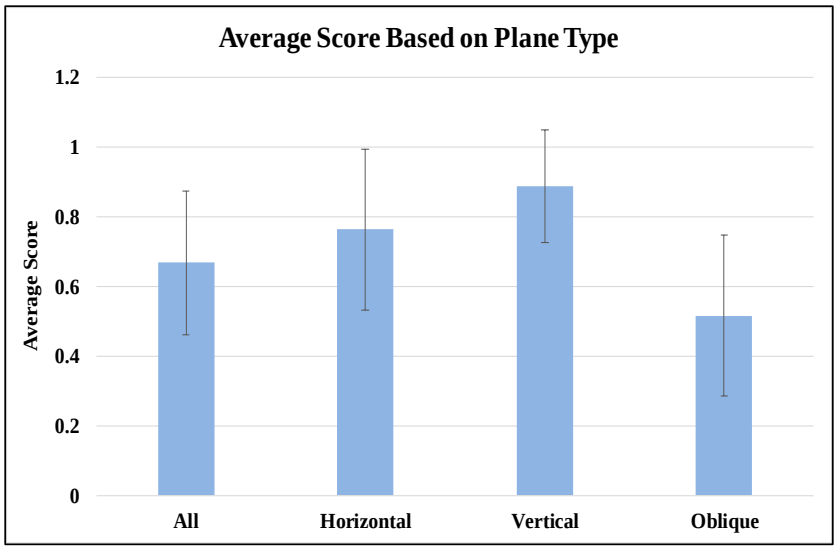


Figure 4.26: Average score based on plane type. Oblique questions are significantly more difficult than Orthogonal questions. For orthogonal planes, vertical questions are easier than horizontal one. This is because viewpoint difficulty for a vertical plane is lower than the horizontal one (see Section 4.2).

4.3.3.2 Patterns of Wrong Answers

For the "patterns of wrong answers", we analyzed the average of the four answer choices (correct, egocentric, alternate, and not possible slice) across the test items of Category 1 and 3 questions. similar to the SBS test, the most frequently chosen wrong answer was the egocentric answer. Not possible and alternate answers were chosen less frequently (See Figure 4.28).

4.3.3.3 Gender Differences

We know the gender of 25 participants (15 male, 10 female). To check the effects of "gender differences" on performance, we conducted a simaple ANOVA. Overall, the average score for male participants was 0.70 (SD = 0.21). The average score for females was 0.62 (SD = 0.16). On average males outperformed females but the difference was not statistically significant ($F(1,23) = 0.67$, $p = 0.42$). Figure 4.29 shows the average

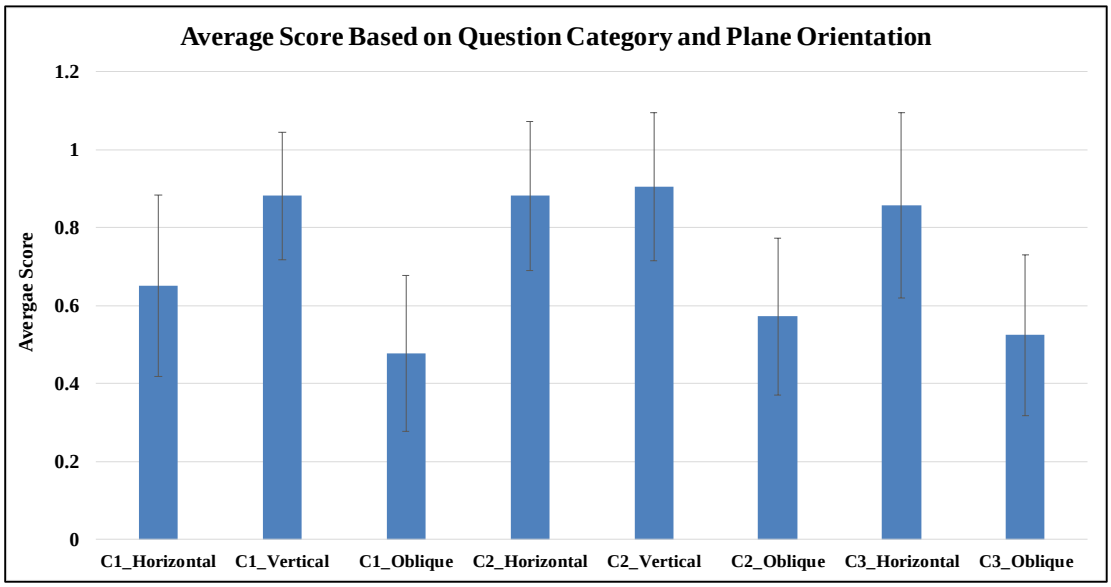


Figure 4.27: Average score based on plane type. Oblique questions are significantly more difficult than Orthogonal questions. For orthogonal planes, vertical questions are easier than horizontal one. This is because the viewpoint difficulty level of a question with a vertical plane is less than the question with a horizontal plane (see Section 4.2).

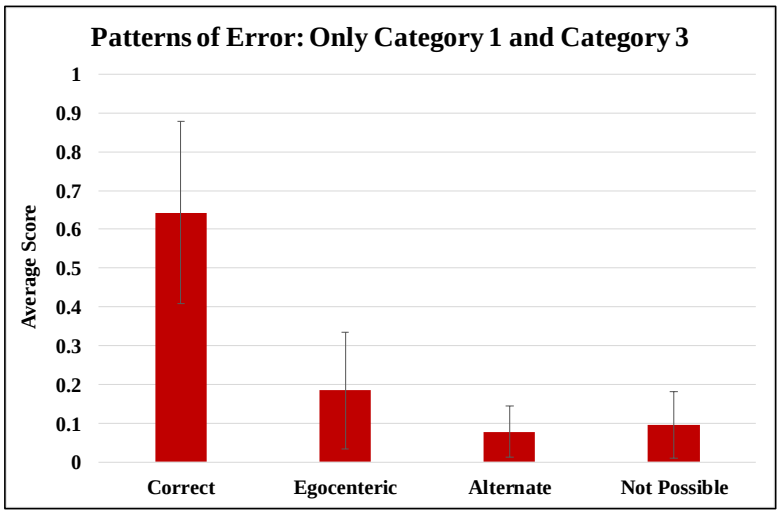


Figure 4.28: Mean proportion of four answer choices on our test versus SBS test.

score for male and female participants.

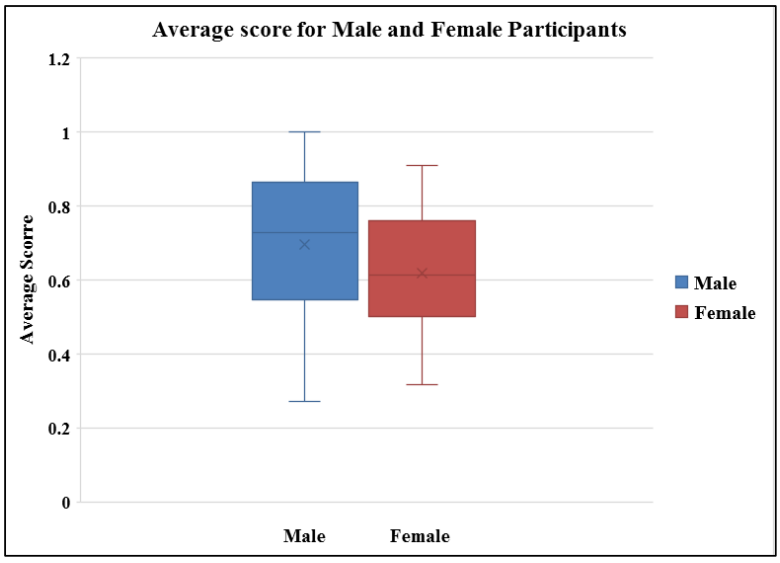


Figure 4.29: Average score for males versus females.

4.3.3.4 Effects of Video Tutorial

To check the "effects of video tutorial", we randomly assigned participants to either watch the video tutorial or read the written instructions (16 participants watched the video tutorial, and 14 participants read the written instructions). The average score of all three question categories for those who watched the video tutorial was slightly higher (M = 0.71, SD = 0.17) than those who just read the written instructions (M = 0.62, SD = 0.19). However, this result is not statistically significant (see Figure 4.30).

4.3.3.5 Background Questionnaire Analysis

Participants indicated that on average they are 24% familiar with the concept of the cross-sections. They ranked their 3D modeling experience to be 17%. The average perceived benefit of the test on enhancing the cross-section understanding was 78%, and the average perceived difficulty of the test was 53%. Finally, participants claimed that

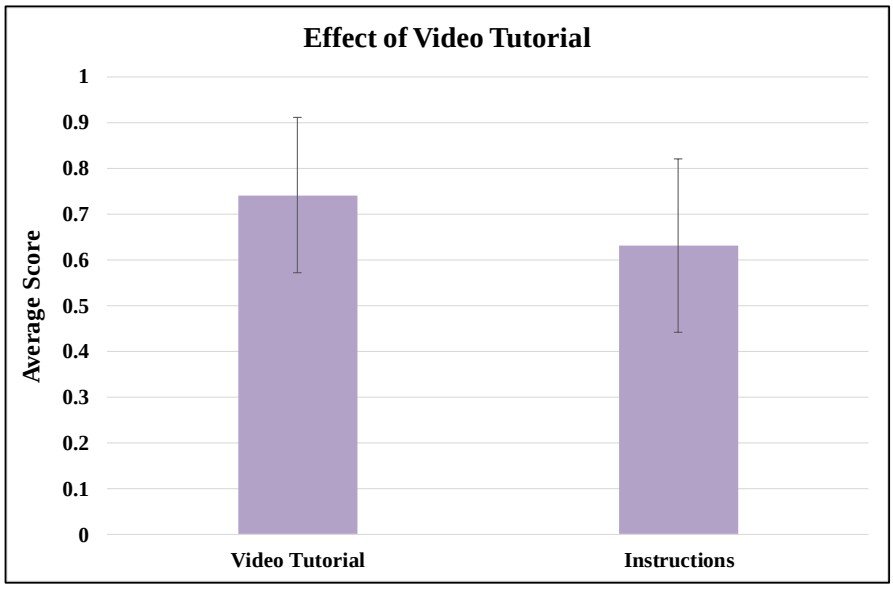


Figure 4.30: Right: Average score (Level 1 and 2) for the the video tutorial versus the written instructions.

they were 63% successful in providing correct answers to the test. Figure 4.31 shows the average of answers to the background question in both Cross-section Understanding Test Version 1 and 2. The level of experience with 3D modeling and familiarity with working with cross-section was lower for participants of Test Version 2. They gave a higher score for test difficulty (found it to be slight more difficult) and lower score for their success in answering the test 2 questions. They also gave a higher score for the perceived benefit of the test.

We also evaluated the relationship between each of the above attributes and the average score of participants. Similar to Test Version 1, there is a non-significant, weak positive correlation between perceived familiarity with cross-sections and average score ($r = 0.45$, $p = 0.13$). Similarly, the correlation between 3D modeling experience and average score is weak and not significant ($r = 0.25$, $p = 0.2607$). Again, there is no significant correlation between average score and both perceived benefit ($r = 0.0808$, $p = 0.72$) and perceived difficulty ($r= 0.2782$, $p = 0.222$). However, there is a significant positive correlation between average score and perceived success ($r = 0.60$, $p < 0.004$).

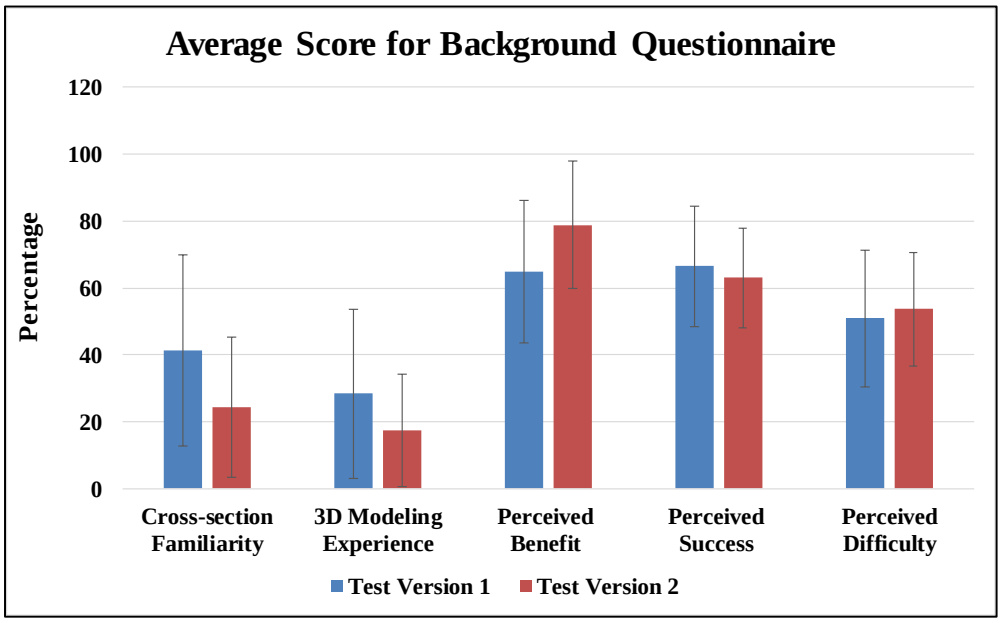


Figure 4.31: Right: Average score (Level 1 and 2) for the the video tutorial versus the written instructions.

4.3.4 Discussion for 2D Cross-section Understanding Test 2

In summary, 2D Cross-Section Understanding Test Version 2 measured participants ability to identify the cross-section of organic shapes.

[**RQ1**] As predicted, performance was better on questions with lower level of difficulty. This result shows that our novel range of difficulty is promising in categorizing test questions based on the different cross-section attributes explained in Section 4.2 .

[**RQ2**] Unlike Test Version 1, for the organic/biological structures we chose shapes that covered a range of difficulty for 2D/3D representations. Most of the objects were bumpy versions of the existing geometric stimuli (e.g., a tubular shape is essentially a cylinder, and a tapering shape is a cone). Again, questions with oblique slicing planes were more difficult than items with orthogonal planes. This result is consistent with the SBS test results, Test version 1, and our defined range of difficulty classification.

[RQ2] The overall pattern of correct and incorrect responses remained the same, indicating that part of the difficulty of the task was visualizing the 3D structure of the objects from the 2D image. Egocentric foil answers are still the most common errors. The fact that the egocentric foil answer was a common error suggests one more time that changing viewpoints relative to the orientation of each structure was still a problem for some of the participants. It is possible that letting the participants interactively move the camera around would further improve their score, or that simply showing two views (instead of one) would be just as useful. This fact is something we should consider in designing our training tool.

[RQ2] Again, similar questions in different categories had different score patterns. This suggests that inverting/modifying the same question (e.g., a question in Category 2 versus Category 1) is not testing the same spatial skill set.

[RQ3] Similar to Test Version 1, study results show that static and video tutorials were both equally effective. This may indicate that the tutorial examples were too trivial as it did not contain any organic shapes (examples were based on geometric structures).

[RQ4] On average male participants outperformed females but the result is not significant. However, we only could analyze the data for 25 participants (5 participants did not answer the background questionnaire). As the average score for males and females of Test Version 2 (males = 0.70, and females = 0.62) is so close to Test Version 1 (male = 0.71, and females = 0.62), we argue that by increasing the number of participants we will see an increase in significance.

[RQ5] On average, comparing to participants of Test Version 1, participants of Test Version 2 rated lower scores for both 3D modeling experience and cross-section familiarity. They also found the test significantly more beneficial in terms of helping them become familiar with the concept of 2D cross-sections. For both participants of the test version 1 and Test Version 2, there is no significant correlation between average test score and the perceived level of 3D modeling experience, cross-section familiarity, benefit, and difficulty. There is a possibility that participants over/under estimated their level of 3D modeling experience and difficulty of the test. However, the significant positive

correlation between average score and perceived success for both tests suggests that participants rated their successfulness more accurately (those who have better performance in the test also rated a higher level of successfulness).

One limitation of the analysis is that, comparing to Test Version 1, we had fewer participants for Test Version 2, and some of the results were not significant. However we could verify our range of difficult based on both test version 1 and Test Version 2 data analysis.

4.4 Conclusions

In this chapter we introduced our 2D Cross-section Understanding Test Version 1, a performance measure for 2D cross-section understanding. Adopted from the Santa Barbara Solids test (SBS test), our 3D stimuli has animations (objectors rotating around y axis), more organic structures (e.g., potato with hole), and two new categories of questions.

Next, we presented a novel range of difficulty for categorizing 2D cross-section understanding tasks with certain difficulty attributes (including viewpoint, mental rotation and 2D/3D representation). Based on this range of difficulty, we created the Test Version 2, with question items of only organic structures that cover different levels of difficulty.

The modified instrument will be used as a pre/post-test to evaluate the effectiveness of the training strategies by measuring participants' 2D cross-section spatial abilities before, and after, the training.

Chapter 5: Developing an Interactive Training Tool for Inferring 2D Cross-sections of 3D Structures

The ability to understand 2D cross-sections of 3D structures plays an important role in many scientific domains including medical imaging, biology, geology, architecture, and engineering [21, 40, 49, 36, 43, 81]. Identifying the correct 2D cross-section of a 3D object requires certain spatial/visualization skills. For example, in medical imaging and particularly 3D volume segmentation, the process is completed (or evaluated) on 2D cross-sections of the 3D data and the segmenter must mentally integrate these into a coherent 3D structure. Therefore, the ability to visualize and infer cross-sections of biological/organic structures is a fundamental skill for segmenters.

The ability to infer a cross section of an object is positively correlated with spatial visualization ability [49, 50, 21, 22]. Unfortunately, not all individuals are equipped with the required spatial skills to successfully understand 2D cross-sections. Therefore, difficulty in understanding cross-sections of 3D structures is an example of how people with low spatial ability might be at a disadvantage in certain scientific fields, including 3D volume segmentation.

There is ample evidence in literature that suggest spatial skills can be improved through experience and training [7, 59, 97, 23, 88, 22, 14, 57, 82, 41]. For example, studies show that playing video games like "Tetris" or first-person shooter (FPS) games has improved performance on spatial ability tests [16, 29, 64, 82, 94]. Still, questions remain about how to best train the spatial skills necessary for inferring 2D cross-sections of 3D objects.

Training novices is challenging because the cross-section understanding by itself is hard and involves many sub-tasks including encoding the spatial characteristics of the structure, imagining slicing the object, removing the section of the sliced object and the cutting plane, and creating an image of the cross-section of the object. Some research has been performed to introduce basic training intervention to help individuals infer the 2D cross-section of simple geometric/geological shapes through drawing and receiving feedback from models [22, 34]. However, these approaches are based on simple virtual

models and/or animations which limit individuals' interaction during the training.

To implement our 2D cross-section training strategy, we introduce a novel interactive tool that uses both geometric and organic 3D models (e.g., structures with branching). Using our tool, participants complete certain tasks with different levels of difficulty. During each task, they learn how to infer 2D cross-sections of 3D objects while interacting with 3D models, observing how a cross-section shape changes after changing the slicing plane location/orientation, and receiving help/feedback if needed. The goal of our training tool is to help novices and people with low spatial skills understand and identify 2D cross-sections of 3D structures. Our main research questions are:

- **RQ1)** Does an interactive training tool effectively help people infer 2D cross-sections of 3D structures?
- **RQ2)** What spatial visualization skills are enhanced after using our interactive training tool for inferring 2D cross-section of 3D structure?
- **RQ3)** What is participants' opinion about the training tool (in terms of perceived success, perceived difficulty, and perceived benefit of the training)?
- **RQ4)** Which training tool features are most effective in enhancing performance?
- **RQ5)** Do gender differences effect performance on 2D cross-section tests/training?

Using controlled studies, we evaluated the effectiveness of our training tool by measuring spatial ability performance of the participants before and after the training.

5.1 Training Tool Design and Development

To deploy our 2D cross-section strategy we introduce a new training tool. This tool consists of multiple tasks with different levels of difficulty for inferring 2D cross-sections, and supports interaction, help, and feedback. Level of difficulty of each task is set based on our novel range of difficulty that was explained in Chapter 4. Development of the training tool involves both task and application design and implementation.

5.1.1 2D Cross-section Strategy and Training Task Design

We designed our training tasks to be goal-oriented and focus on skills needed to understand the relationship between cross-section creating and viewpoint direction, slicing plane orientation, and 3D structure. The tasks involve identifying cross-sections based on some certain characteristics of the 3D structure, locally adjusting the plane to create a certain change in the 2D target cross-section, and placing the plane to cut the 3D structure. Designing the tasks was an iterative process and we developed many design alternatives. Tables 5.1, 5.2, and 5.3 show the final version of our training task design along with the sketches of the 3D structures. In our final design, the training tool has three levels:

Level 1: Consists of three tasks with simple 3D objects (hourglass and tapering structures). The goal of level 1 tasks is to train participants to identify the cross-sections of 3D objects through simple plane movements/rotations.

- *Task 1: "Move the plane so that it creates a cross-section for the skinniest/ thinnest part of the given 3D shape".*

This task involves a simple plane movement. Plane orientation is orthogonal, and the 3D structure is a simple hourglass shape. The viewpoint with respect to the 3D structure is orthogonal. To accomplish this task, move the plane to the skinniest part, cut the shape and rotate object/change viewpoint to see the correct cross-section.

- *Task 2: "Move the plane so that it creates a cross-section for the fattest/ thickest part of the given 3D shape".*

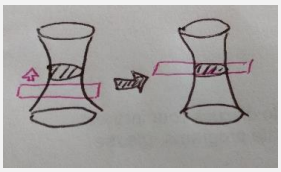
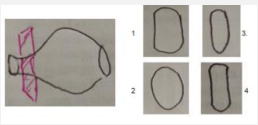
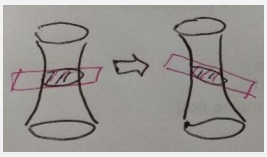
This task is similar to Task 1. It involves a simple plane movement. Plane orientation (with respect to the object) is orthogonal, and the 3D structure is a simple tapering shape. However, the viewpoint with respect to both the plane and the 3D object is not orthogonal (the objects are not facing the initial viewpoint). To accomplish this task, mentally move the plane to the thickest part, cut the shape and appropriately rotate object/change viewpoint to see the correct cross-section.

- *Task 3: "Adjust the plane to change the cross-section shape from a circle to an oval".*

This task involves a simple plane rotation. Plane orientation is initially orthogonal (horizontal), and the 3D structure is a simple hourglass shape. The viewpoint with respect to the 3D object is orthogonal. To accomplish this task, mentally rotate the plane to change its orientation from orthogonal into oblique, cut the shape and rotate object/change viewpoint to see the correct cross-section.

Table 5.1 shows Level 1 tasks.

Table 5.1: Level 1 Tasks.

Task #	Definition and Outcome	Objects	Level of Difficulty	Figure
1.1	**Adjust the plane so that it creates a cross-section for the skinniest part of the given 3D shape** **Outcome:** Build an understanding of cross-section changes of simple structures by simple plane transition/rotation	-hourglass -horizontal Plane	1	
1.2	**Adjust the plane so that it creates a cross-section for the fattest part of the given 3D shape** **Outcome:** Build an understanding of cross-section changes of simple structures by simple plane translation/rotation	-Tapering Structure - Vertical Plane	2	
1.3	**Adjust the plane to change cross-section from circle to oval** **Outcome:** Build an understanding of cross-section changes of simple structures by simple plane translation/rotation	-hourglass -horizontal Plane	3	

Level 2: Consists of two tasks with a complex 3D object (Branching Y shape). The goal of level 2 tasks is to train participants to understand the cross-sections of a more complex organic 3D object while implementing plane movements/rotations.

- *Task 1: "Adjust the plane so that it creates a cross-section that cuts across both branches but not the stem."*

This task involves plane movements. Plane orientation is orthogonal. However, the

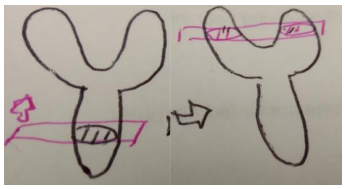
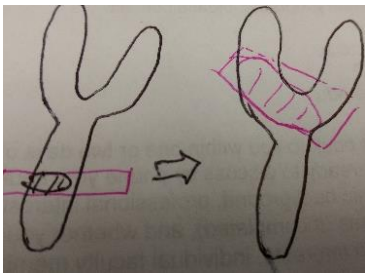
viewpoint with respect to the 3D object is not orthogonal. To accomplish this task, move the plane above the place where the branching starts, then cut the shape and rotate object/change viewpoint to see the correct cross-section.

- *Task 2: "Adjust the plane to create a single, oval-ish cross-section that crosses both one branch and the trunk."*

This task involves both plane movement and rotation. Plane orientation (with respect to the object) is orthogonal. However, the viewpoint with respect to both the plane and the 3D object is not orthogonal. To accomplish this task, move the plane to the branching point, change its orientation to oblique, cut the shape and appropriately rotate object/change viewpoint to see the correct cross-section.

Table 5.2 shows Level 2 tasks.

Table 5.2: Level 2 Tasks.

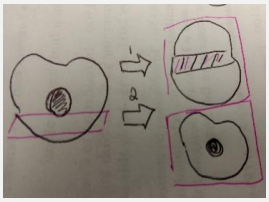
Task #	Definition and Outcome	Objects	Level of Difficulty	Figure
2.1	**Adjust the plane so that it creates a cross-section that includes both branches.** **Outcome:** See cross-section changes for an organic shape by simple translations	Branching Structure	**4**	
2.2	**Adjust the plane to create an oval cross-section that captures both one branch and the trunk.** **Outcome:** See cross-section changes for an organic shape by simple rotations and transtions	Branching Structure	**5**	

Level 3: Consists of one task with a more complex 3D object (potato shape with hole). The goal of level 3 ask is to train participants to understand the cross-sections of an organic 3D object while implementing multiple plane movements/rotations and viewpoint changes.

- *Task: "Adjust the plane so that the hole in the object creates a circular cross section (surrounded by an oval-ish cross section for the outside of the shape). Place the plane through the fattest part of the shape."*

This task also involves both plane movement and rotation. Plane orientation (with respect to the object) is initially orthogonal. However, the viewpoint with respect to both the plane and the 3D object is not orthogonal, therefore it is not possible to see the hole from the initial viewpoint. To accomplish this task, change the viewpoint to see the hole, move the plane to the center of the potato shape, change its orientation to fully capture the hole, cut the shape and appropriately rotate object/change viewpoint to see the correct cross-section. Table 5.3 shows Level 3 task.

Table 5.3: Level 3 Task.

Task #	Definition and Outcome	Objects	Level of Difficulty	Figure
3.1	**Adjust the plane to capture center of the hole to create a cross-section with a circle hole inside it.** **Outcome:** See cross-section changes for an organic shape by combination of translations and rotations	Potato With Hole	6	

As the level of tasks increases, the tasks become more difficult. Starting from simple objects with just one simple plane movement or rotation, at the end we have a task with more complex organic objects, that requires mentally changing viewpoints, and rotating/moving the plane multiple times.

5.1.2 Training Tool Application Design

We instantiated the task design in our training tool application. We developed the training application to look like a 3D game with multiple features, focusing on simulating spatial/visualization skills needed for understanding 2D cross-section of a 3D structure (e.g., plane rotation/movement). By embodying the task design principles described in the previous section, we insured that our training tool covered different spatial tasks

with a range of difficulty.

To evaluate that our training tool design worked well in practice, we employed a participatory design methodology [86]. For each design iteration we ensured that the training tool prototype embodied the tasks design principles. We then asked an end user to review the design and try to accomplish multiple tasks with the prototype. The initial iterations involved low-fidelity sketches and paper-based prototypes to encourage rich participant feedback. Figure 5.1 a) shows an early prototype sketch in the design process.

The challenge for evaluating the low-fidelity prototype was that the tasks involved interactions (such as moving the plane), which was not fully possible using the low-fidelity prototype. However, we successfully elicited major suggestions for usability refinements, such as the appearance of the main window, visibility and consistency of buttons, providing feedback, and adding help options. After three iterations we implemented a higher-fidelity prototype in myBalsamiq [8]. Figure 5.1 b) shows a page example of the prototype. We conducted additional usability tests with our higher-fidelity prototype and received more usability feedback about the representation of buttons and sliders, and how to make them simpler and more meaningful so that they show the possible actions user can take with them (enhancing UI affordance [63]).

Using our prototype, we then implemented our training tool in Unity [96]. Unity is a cross-platform game engine which is primarily used to develop both 3D and 2D games and simulations. We used both the 2D and 3D tools and UI features of unity to create the game environment. Most of the training tool scripts are in C#, but we also used materials, shaders, and textures to embody our 3D models in the game. Inspired by the work prseneted in [3], we developed our customized shaders to create cross-sections of 3D structures. These shaders divided the space into two partitions, the one under the plane was rendered, and the partition above the plane was discarded.

We conducted three additional pilot tests with our training tool, and as expected, the feedback about our application focused on more minor behavioral adjustments. For each test, an end user was given a short tutorial on the tool user interface, then asked to complete the required tasks shown in the tool. A researcher observed each test and took notes about how participants used the application. We revised the application after each test to resolve participants' problems. Figure 5.1 c) shows the final main screen for one of the training tasks implemented in Unity. Below we describe each of the features

provided by the tool.

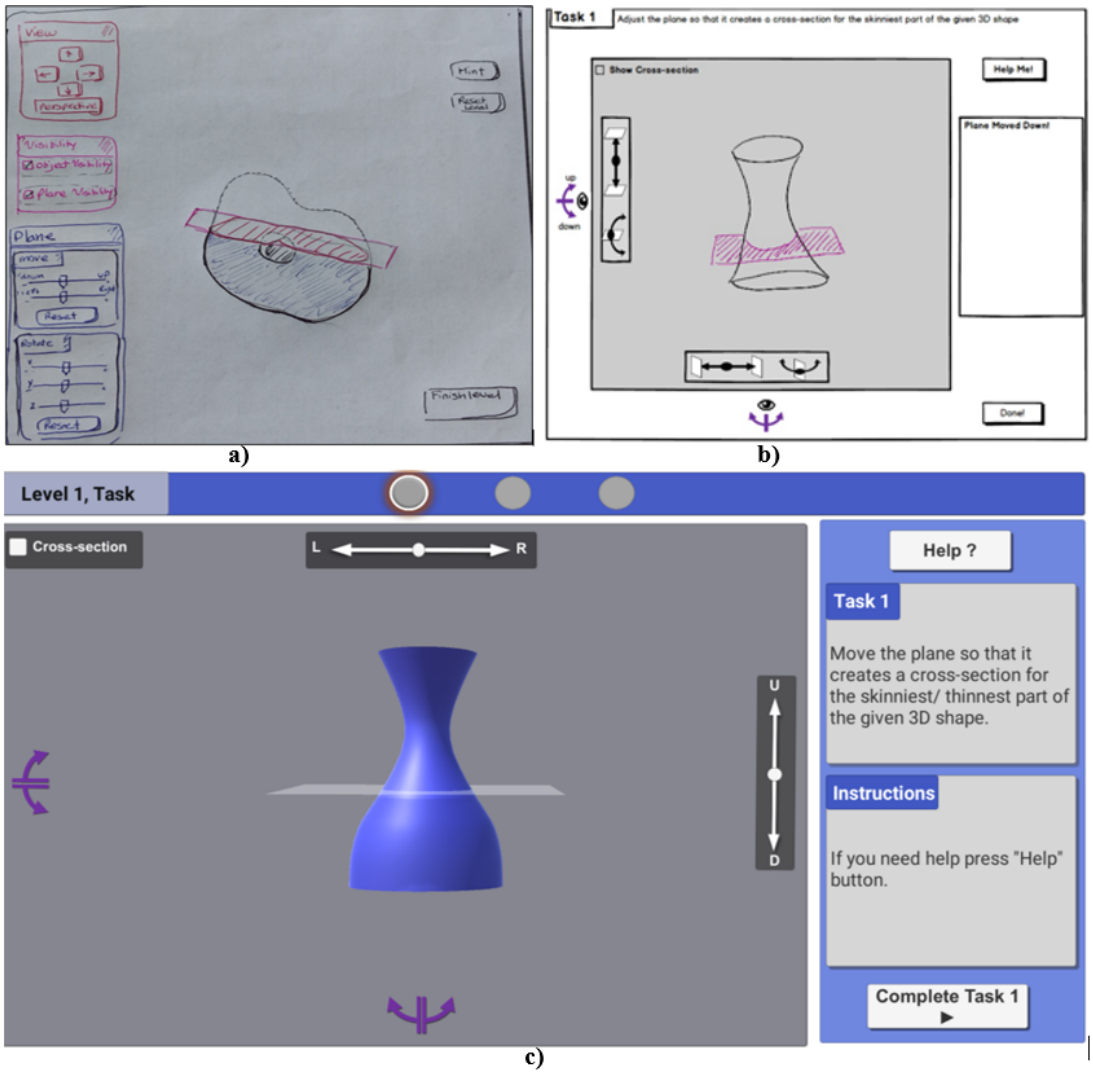


Figure 5.1: Sketches and prototypes: a) A low-fidelity sketch example; b) Mybalsamiq prototype; c) Unity Implementation.

5.1.2.1 Levels and Tasks

The training tool has three levels and each level has at least one task (level 1 has three, level 3 has two, and level 3 has one task). Each task consists of two modes: "Play" and "Solution". Figure 5.2 shows these two modes for one of the tasks.

The UI for each mode has three separate panels: 1) Header panel: which has the level and task title along with a progress status image. 2) Main panel: which has the 3D model along with the buttons and sliders to move/rotate the plane, change viewpoint, and see cross-section changes. 3) Help and answer panel: the panel appears differently in the two UI modes. In the "Play" mode, there is a help option button, a sub-panel to show the task, a second sub-panel to show the help instructions, and a "Complete Task" button. In the "Solution" mode UI, there is a "Show Answer" button, two sub-panels to show the task and answer guides, and "Go to Next Task" or "Finish Level" buttons.

In the "Play" mode, participants complete a given task (for example in task1 move the plane to the skinniest part of the 3D model). By hitting the "Complete Task" button they confirm that they have completed the task, and then would be directed to the "Solution" mode to see an answer for creating a correct cross-section. By clicking the "Next" button they are either directed to the next task of the same level or to the next level. In total we have six "Solution" mode pages. There are also transition pages between levels to make sure participants understand that the level of the training has changed.

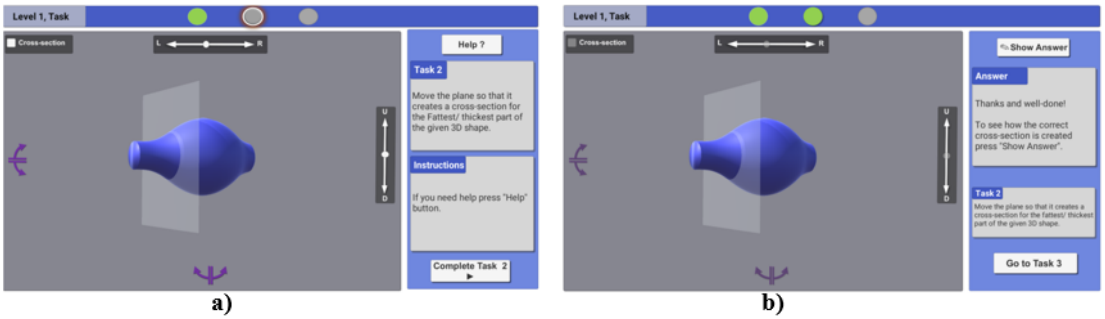


Figure 5.2: Training tool UI modes: a) Play mode; b) Solution mode.

5.1.2.2 3D Models

We created four different 3D structures using 3D Studio Max, and ZBrush. We then imported the object files (*.obj) of the 3D models into the Unity and added materials, textures, and shaders to these objects to include shadows, colors, light, and cross-section effects. Level 1 uses the hourglass and tapering shape 3D models. Level 2 uses the branching structure, and Level 3 uses the potato with a hole 3D model. Figure 5.2 shows our 3D models in Unity.

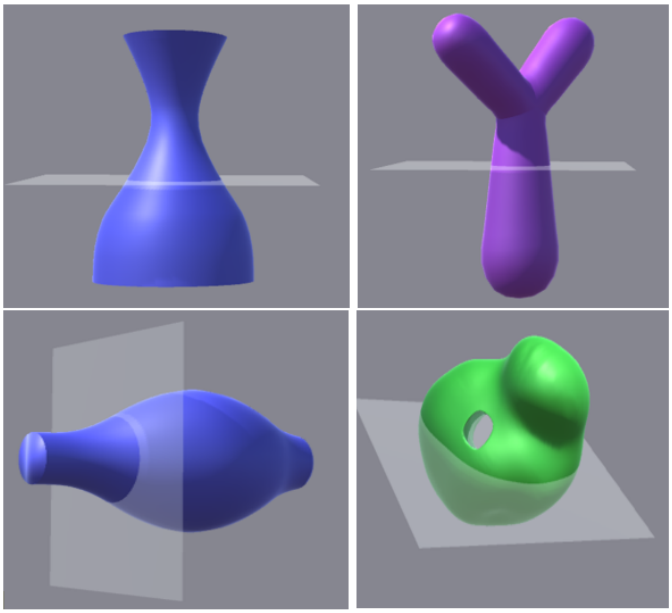


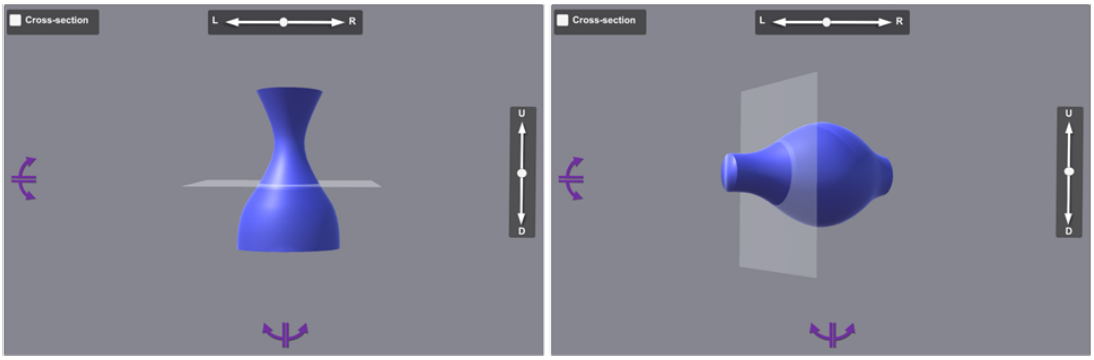
Figure 5.3: 3D models for the training

These 3D objects cover a range of difficulty for the "3D Object Representation" category (see Section 4.2). The hourglass and tapering shape are simple symmetric objects, the branching shape is a simple asymmetric organic object, and the potato with a hole is a nested, asymmetric organic shape.

5.1.2.3 Viewpoint Visualization

Another spatial skill necessary for 2D cross-section understanding is visualizing the correct viewpoint from which an object is observed. In our training tool, the initial view-

point with regards to the objects (3D model and slicing plane) has different orientations (orthogonal or non-orthogonal). By combining the shape complexity of 3D models with the viewpoint orientation, we have a range of difficulty for viewpoint visualization. For example, in Level 1 Task 1, the initial viewpoint with regards to the 3D model (an hourglass shape) is orthogonal, while in Level 1 Task 2, the viewpoint with regards to the objects (both the 3D model and the plane) is not orthogonal, making the task more difficult (see Figure 5.4).



b) Level 1 Task 1: Orthogonal Viewpoint a) Level 1 Task 2: Oblique Viewpoint

Figure 5.4: Two Viewpoints. a) Viewpoint with regards to the 3D model is orthogonal. b) The viewpoint with regards to both the 3D model and the plane is not orthogonal

We added UI buttons to allow participants to change the views and see the object or plane. The representation and direction of these buttons are designed in a way to show those possible actions participants can take with them. The "Up" or "Down" buttons (labeled 1 in Figure 5.5) allow to observe the objects from top view. The "Left" or "Right" (labeled 2 in Figure 5.5) buttons allow participants to rotate the camera around the object (same as rotating the object around y axis).

Previous research claims that full viewpoint interactivity can possibly confuse participants [50, 52]. Therefore, we have limited both the interaction and number of possible viewpoints. For example, participants are not able to change the views or rotate the camera (objects) freely just by mouse clicking. Instead, they must click the "Up"/"Down"/"Left"/"Right" buttons to change the view/rotate the camera.

Observing the objects from different views is helpful for completing some of the 2D cross-section inferring tasks. For example, in the Level 3 Task, we have a 3D potato shape

with a hole. However, the hole is not visible in the initial viewpoint. The necessary first step for this task is to rotate the object (change view to left or right) to see the hole (see Figure 5.5).

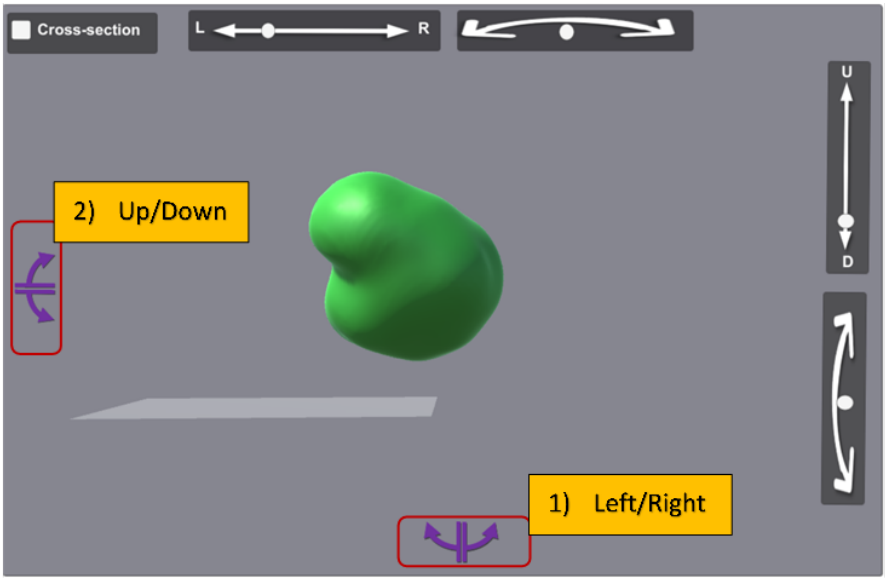


Figure 5.5: Level 3 Task. The hole is not visible in the initial viewpoint. Using "Left" or "Right" buttons (labeled 2) we can rotate the object and change the view to see the hole.

5.1.2.4 Plane Movement and Rotation

One of the skills needed for 2D cross-section understanding is to mentally adjust (rotate/move) the plane. To simulate this skill in our training tool, we added features to allow participants to move and rotate the plane. We included four sliders as UI elements for the plane movement/rotation. The representation and direction of "movement sliders" (straight) are different from the "rotation sliders" (curved) to show possible actions participants can take with them. Using the "movement sliders", participants can move the plane up/down, back/forth, and left/right. Using the "rotation sliders" participants can rotate the plane in different orientations.

The training tool tasks cover the range of difficulty for "Mental Rotation/movement".

For example, for Level 1 Task 1, participants only need to move the plane up to the skinniest part, while Level 2 Task 2 requires more complex mental movements/rotations of the plane. The UI sliders help participants to simulate the mental rotation/movement appropriately. Not all the sliders will be visible in each task. In Level 1 tasks, that involve simple plane movements or rotations, only the movement sliders or the rotation sliders are visible. For the Level 2 task, which requires multiple plane movements and rotations, all sliders are visible. Figure 5.6 shows the Level 2 task along with all four sliders: "Movement sliders" (straight arrows labeled 1 and 2) and "rotation sliders" (curved arrows labeled 3 and 4)

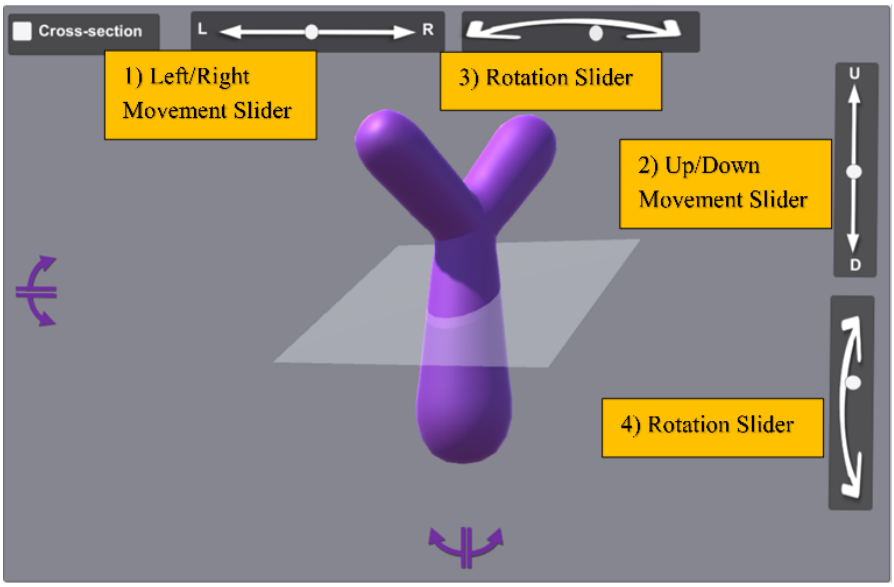


Figure 5.6: Level 2 Task 2 with all four sliders: "movement sliders" (straight arrows labeled 1 and 2) and "rotation sliders" (curved arrows labeled 3 and 4)

5.1.2.5 Showing Cross-section from the Cut-away

To identify a correct 2D cross-section, one should mentally cut the 3D structure and imagine the 2D representation from that cut-away. This could be a difficult task depending on the shape complexity of the 3D structure and the orientation of the slicing plane. In our tool, participants can see the cross-section by checking the "Cross-section"

checkbox. Upon selecting the checkbox, the 3D structure is cut from the location of the slicing plane. The 2D cross-section is then visible with a different black and white texture. To fully observe the cross-section, it might be necessary to change the view. Figure 5.7 shows the cross-section for the skinniest part of the 3D model in Task 1 Level 1, after selecting the checkbox and changing the view to see the model from the top.

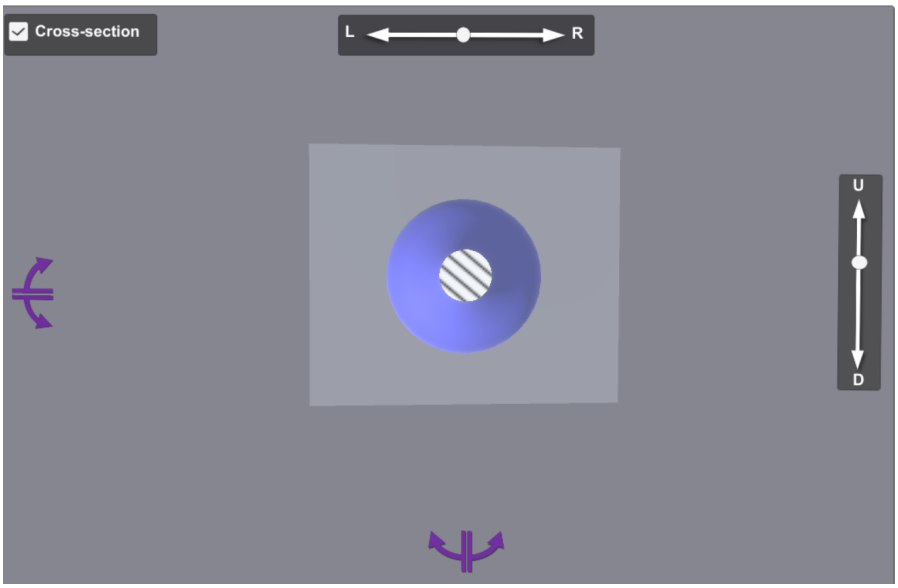


Figure 5.7: Task 1 Level 1 (hourglass shape), after selecting the checkbox and changing the view to see the model from top. The correct cross-section is a small circle shown with a black and white texture.

5.1.2.6 Help and Feedback

To help participants accomplish the tasks we provided on-demand help options. By hitting the "Help?" button, the help instructions are shown in the UI (see Figure 5.2 a.) Different "Help" instructions are shown to users based on interaction factors such as number of times the "Help?" button has been clicked, plane location/direction, and view orientation.

5.1.2.7 Solution

One of the important parts in the training is to show participants how the correct cross-section for each task is created. As mentioned previously, the answers are implemented in the "Solution" mode. By clicking the "Show Answer" button, participants can see how the correct cross-section is created. Then the actions needed to correctly create a cross-section will be shown to them step by step. See Figure 5.3 b) for the "Solution" mode. In this mode, some of the buttons and sliders are disabled to limit the participants' interaction with the tool (we just want our participants to see the solution and to not change anything in the UI). However, participants are free to click and see the answer as many times as they want.

5.1.2.8 Score and Task Evaluation

To evaluate participants' performance during the training, we logged elapsed time and gained score for each task. If participants successfully complete a task, they will earn 100 points. The maximum score is 600 and there is no penalty for wrong answers. "Finish time" of each task and the score is stored in a .txt file. We also automatically screen capture the latest state of each participant's work upon completing a task, and save it for future analysis.

5.1.2.9 Action Logging

Every action that the participants take during the training session is logged by the tool application in a .csv format file. The actions include, but are not limited to, "Help?" button clicks, "Show Answer" button clicks, changing views, working with sliders, seeing the cross-section, and finishing a task. We use the results of these logs for our future analysis (e.g., how many times a participant requested help). Figure 5.8 shows part of a log file for one participant.

5.2 Study Methodology

To evaluate our training tool, we conducted user studies. In this section we fully describe our methodology to investigate our training tool's effectiveness with end users.

```
2018/03/19 16:44:34, Slider_Left_Right_Changed, Level 1 Task 2 Slider Left_Right is changed.
2018/03/19 16:44:34, Slider_Left_Right_Changed, Level 1 Task 2 Slider Left_Right is changed.
2018/03/19 16:44:34, Slider_Left_Right_Changed, Level 1 Task 2 Slider Left_Right is changed.
2018/03/19 16:44:34, Slider_Left_Right_Changed, Level 1 Task 2 Slider Left_Right is changed.
2018/03/19 16:44:34, Slider_Left_Right_Changed, Level 1 Task 2 Slider Left_Right is changed.
2018/03/19 16:44:36, Help_Button_Pressed, Level 1 Task 2 Help Button is pressed.
2018/03/19 16:44:36, Check_CrossSection, Level 1 Task 2 Check_CrossSection is checked.
2018/03/19 16:44:37, Down_Button_Pressed, Level 1 Task 2 Down_Button is pressed.
2018/03/19 16:44:38, Up_Button_Pressed, Level 1 Task 2 Up_Button is pressed.
2018/03/19 16:44:39, CompleteTask_Button_Pressed, Level 1 Task 2 Complete Button is pressed.
2018/03/19 16:44:40, GoNextTask_Button_Pressed, Level 1 Task 2 Evaluation Go to Task 3 Button is pressed.
2018/03/19 16:44:41, CompleteTask_Button_Pressed, Level 1 Task 3 Complete Button is pressed.
2018/03/19 16:44:42, FinishLevel_Button_Pressed, Level 1 Task 3 Evaluation Finish Level Button is pressed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:47, Slider_Left_Right_Changed, Level 2 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:48, Help_Button_Pressed, Level 2 Task 1 Help Button is pressed.
2018/03/19 16:44:49, CompleteTask_Button_Pressed, Level 2 Task 1 Complete Button is pressed.
2018/03/19 16:44:50, GoNextTask_Button_Pressed, Level 2 Task 1 Evaluation Go to Task 2 Button is pressed.
2018/03/19 16:44:51, Slider_YRotate_Changed, Level 2 Task 2 Slider YRotate is changed.
2018/03/19 16:44:51, Slider_YRotate_Changed, Level 2 Task 2 Slider YRotate is changed.
2018/03/19 16:44:51, Slider_YRotate_Changed, Level 2 Task 2 Slider YRotate is changed.
2018/03/19 16:44:52, CompleteTask_Button_Pressed, Level 2 Task 2 Complete Button is pressed.
2018/03/19 16:44:54, ShowAnswer_Button_Pressed, Level 2 Task 2 Evaluation Show Answer Button is pressed.
2018/03/19 16:44:55, FinishLevel_Button_Pressed, Level 2 Task 2 Evaluation Finish Level Button is pressed.
2018/03/19 16:44:58, Slider_Left_Right_Changed, Level 3 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:58, Slider_Left_Right_Changed, Level 3 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:58, Slider_Left_Right_Changed, Level 3 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:58, Slider_Left_Right_Changed, Level 3 Task 1 Slider Left_Right is changed.
2018/03/19 16:44:58, Slider_Left_Right_Changed, Level 3 Task 1 Slider Left_Right is changed.
```

Figure 5.8: Actions logged by the training tool application.

5.2.1 Participants and Experiment Design

We used a between-subject study design to evaluate our training tool. The treatment group used our training tool while the control group played the game Word Whomp [5] instead. Word Whomp was proven to be an appropriate game for the control group while evaluating the effectiveness of 3D games on enhancing 3D spatial abilities [82].

We recruited 60 participants (33 males, 25 females, and 2 did not declared their gender) from the local community and university.

The experiment session was conducted in groups of up to 5 participants at a time. We assigned each participant to the first available experiment session that met their time constraints and then randomly assigned each session to either the control or treatment condition. A total of 30 participants were assigned to the control (Game) and 30 participants took part in the treatment (Training) group. Each study session consisted of three parts: pre-test, main training or game, and post-test.

5.2.2 Materials

5.2.2.1 Spatial Ability Performance Measures and Tests

We used the below tests to measure spatial ability performance of our participants before and after the training:

Mental Rotation Performance: We measured 3D mental rotation performance by using an online version (implemented in Qualtrics) of the redrawn Vandenberg and Kuse [100] Mental Rotations Test (MRT) by [68]. The MRT consists of 24 questions and is divided into two sets of 12 items. Each question consists of five 3D block figures (see Figure 5.9). The target block figure on the top has to be compared to four similar blocks below. In each item, two of the four figures are rotated versions of the target (correct alternatives), whereas the other two are distractor figures. The task is to detect the two correct alternatives. 2 points is given if both correct figures are chosen, 1 point if one correct answer is chosen and 0 points if no correct answer is chosen, therefore the maximum score for each part of this test is 24. We used the first set in our pre-test and the second set in our post-test. Each set takes 8 minutes.

Card Rotations (S-1): This test is used to evaluate 2D spatial relation performance [33], and has two parts. Each part consists of a set of 10 items. Each item has nine 2D figures with the target shape at the left. Participants decide whether each of the eight other possible options represent either a rotation (choose "S" to confirm the option is the same as target) or a mirror image of the target (choose "D" to confirm the option is different from the target). Every correct answer is worth one point. Therefore, the maximum score for each part of the test is 80. Each set takes 5 minutes. The time limit for the original test is 3 minutes. Our pilot studies showed that 3 minutes is not enough to complete the online version of the test. We consider 2 extra minutes to overcome this problem. We implemented this test in Qualtrics (see Figure 5.10 for an example item). We used part 1 in our pre-test and part 2 in the post-test.

Viewpoint Visualization Performance: We used a modified version of Guay's Visualization of Views Test [27] to measure the viewpoint visualization ability. In the test, participants need to tell which viewing position a picture of a 3D object was taken from

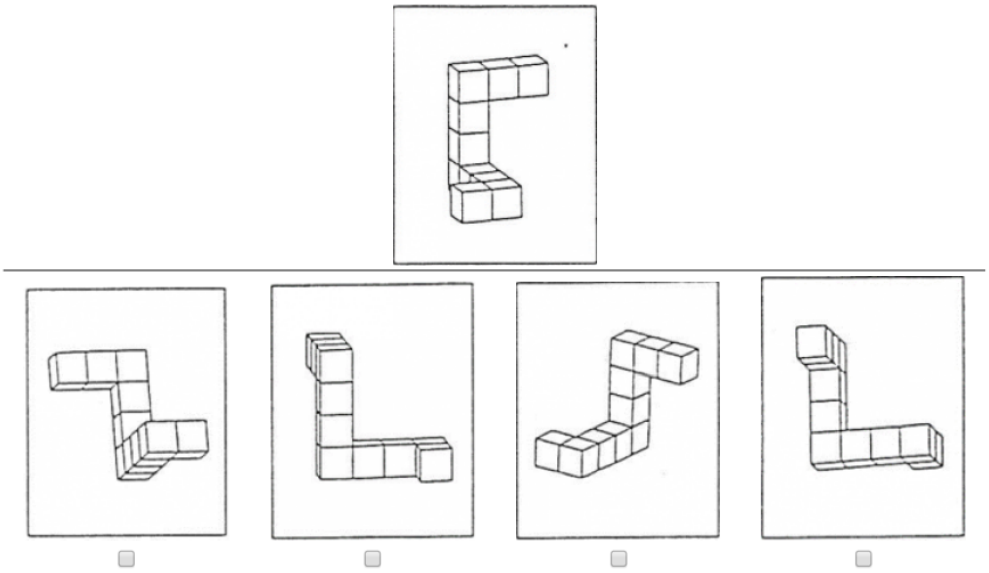


Figure 5.9: An example item from Mental Rotation Test (MRT) [100].

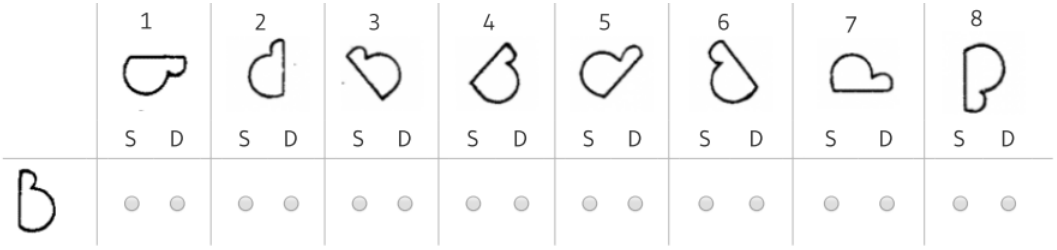


Figure 5.10: An example item from Card rotation S-1 test [33].

(there are seven alternatives). This test is broken into two sets of 12 questions. The test lasts for 8 minutes. The score is the number of items answered correctly, and the maximum score for each set is 12. Figure 5.11 shows a sample of the Guay's Visualization of Viewpoints test implemented in Qualtrics.

2D Cross-section Understanding Performance: We used our 2D cross-section test version 2 (explained in 4.3) to measure the performance of our participants in understanding 2D cross-section of different 3D organic structures. The test has 26 questions.

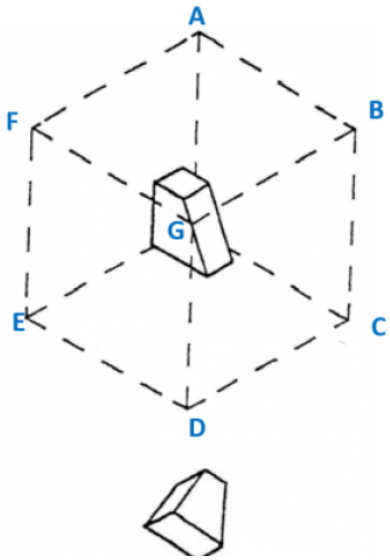


Figure 5.11: An example item from Guay's Visualization of Viewpoints test [27].

There are 12 questions of Type 1, six questions of Type 2, four questions of Type 3, and four controlling questions to check consistency. We used 13 questions for the pre-test (see Appendix item "2D Cross-section Understating Test Version 2, Part 1") and 13 questions for the post-test (see Appendix item "2D Cross-section Understating Test Version 2, Part 2"). The questions in each set were chosen in a way to make sure: 1) the cumulative score of difficulty for pre and post sets are the same (cumulative difficulty grade of each set=41); and 2) The average score of our MT participants (see Section 4.3) for both sets are nearly equal (average score for each set=0.66). The maximum score for each part of this test is 13. Figure 5.12 shows three examples of each question type.

5.2.2.2 Training and Game

The treatment group received our training strategy, while the control group played the game. As explained in the previous sections, the training tool is implemented in unity and consists of six tasks in three different levels. The control group played the game Word Whomp, in which participants are asked to make words out of a group of six random letters to earn points (see Figure 5.13).

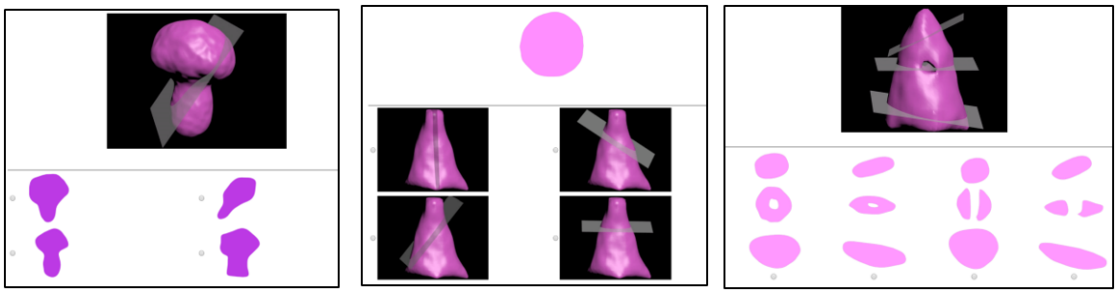


Figure 5.12: Our test instrument for inferring 2D cross-sections (see Chpater 4, Section 4.3): Three examples of each question type.

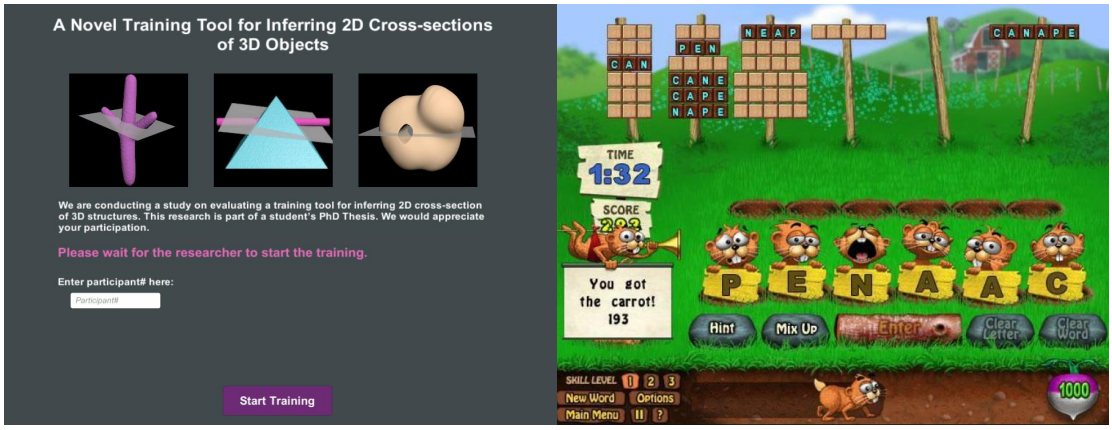


Figure 5.13: Left: First page of the training tool. Right: Word Whomp, EA Games.

5.2.2.3 Background and Post-test Questionnaire

We created a questionnaire to gather more information about our participants. The information we are interested in includes age, gender, level of education, level of experience working with 3D models and cross-sections, perceived benefit of the training, and a self-evaluation of success.

5.2.3 Procedure

Each experiment consists of three sub-sessions: pre-test, main training or game, and post-test. In the pre-test, all 30 participants completed the first part of our 2D cross-

section understanding test. Our pilot studies showed that participants became extremely tired if they completed all the other three tests. Therefore, to avoid participants' fatigue, 10 of them were randomly assigned to complete one of the other tests: Mental Rotation Test (MRT); Visualization of View Test(VV); or S-1. Table 5.4 shows the participants distribution based on gender and test group.

Table 5.4: Number of participants in each test group categorized based on gender.

Group	**Training**				**Game**			
Test Measure	**2D CS**	**MRT**	**VV**	**S1**	**2D CS**	**MRT**	**VV**	**S1**
Male	16	6	5	5	17	6	5	6
Female	14	4	5	5	11	2	5	4

The treatment group (also called training group) went through the 30 minutes of training while the control (game) group engaged in 30 minutes of playing the game.

During the training, a researcher introduced participants to the training tool application via a brief hands-on tutorial (see Appendix item "Training Tool Tutorial") that explained how to use it. Participants then had three minutes to explore the tool on their own. To avoid learning effects, the tutorial involved a task that was different from the main training tasks: "Learn to work with different features of the training tool" (see Figure 5.14 for the training practice example). The main training tasks followed the practice session. We asked our participants to complete all those six tasks in the three levels as best as they could and spend some time on the "Solution" parts to understand how the cross-sections were created. Our application logged all participant interactions (e.g. button clicks) and computed their scores. We used this logged data for future analysis. However, the score was not shared with our participants.

The post-test followed the main training/game session. We asked all of our participants to complete the second part of the 2D cross-section understanding test, and then the participants completed one of these corresponding tests (MRT, VV, and S1). At the end of the post-test session, we collected demographic data from participants through our background questionnaire. This included information such as age, gender, and familiarity with the topics for the experimental task. To understand participants' reactions to our training tool, the post-test questionnaire also asked participants about the tool and their perceived benefit from the training. We asked participants to skip the questions rather than provide incorrect information if they did not wish to answer them.

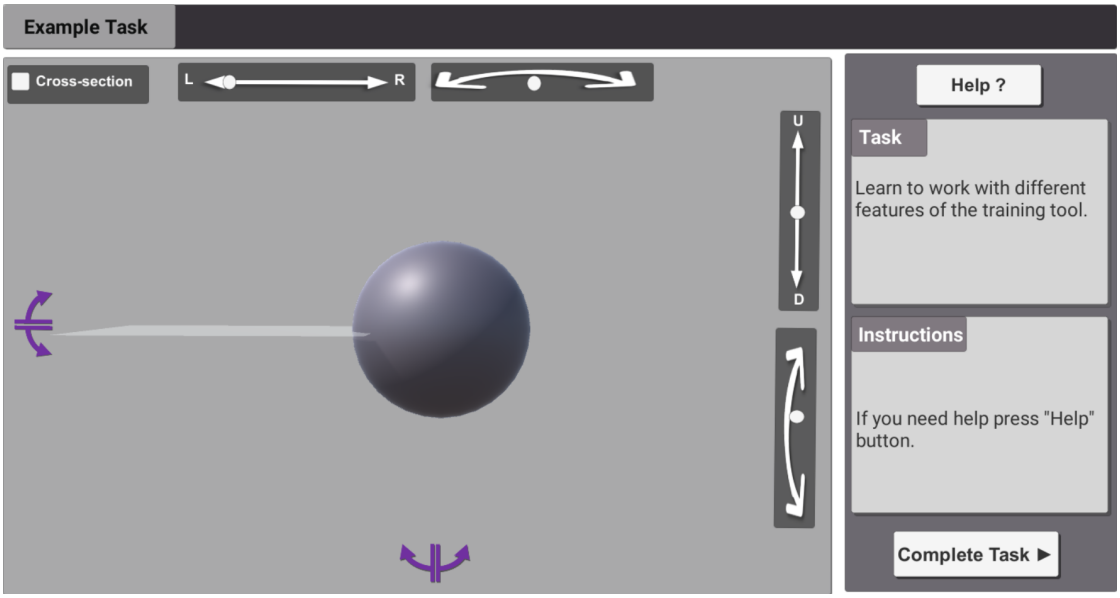


Figure 5.14: Training tool tutorial task

The entire study session took no longer than two hours.

5.3 Results

5.3.1 Effects of Training Tool on Performance

Table 5.5 shows descriptive statistics for all the test measures (pre and post), by test groups (training or game). As visible in the table, there were no significant differences in performance between the two test groups on any of the pre-test measures (all ts $<$ 0.75, and all p-values $>$ 0.31). As predicted, performance on the cross-section test was significantly correlated with MRT ($r = 0.63$, $p = 0.0029$),and VV ($r = 0.59$, $p = 0.0048$), but not S-1 ($r = 0.43$, $p = 0.06$).

Analyses of within-subject improvement from pre- to post test (paired t-test) for each of the test measures are:

- **2D Cross-Section Test:**
  - Training (pre and post): $t = 7.30$, $p < 0.0001$; result is extremely statistically

Table 5.5: Mean (SD) performance for each test measure by test group.

Test Measure	Training		Game	
	Pre	Post	Pre	Post
2D Cross-section	9.1 (2.03)	11.6 (1.53)	9.6 (1.7)	9.53 (2.4)
MRT	17.3 (3.74)	19.8 (3.34)	17.7 (3.49)	16.6 (4.7)
VV	5.2 (3.09)	7.5 (2.83)	5.9 (2.96)	6.2 (1.91)
S1	71.8 (5.11)	73.6 (3.23)	72.55 (10.94)	69.22 (14.61)

significant.

- Game (pre and post): t = 0.15, p = 0.88; result is not statistically significant.

- **Mental Rotation Test (MRT):**
  - Training (pre and post): t = 3.55, p = 0.0062; result is very statistically **significant**.
  - Game (pre and post): t = 0.56, p = 0.58; result is not statistically significant.

- **Visualization of Views Test (VV):**
  - Training (pre and post): t = 4.44, p = 0.0016; result is very statistically **significant**.
  - Game (pre and post): t = 0.43, p = 0.67; result is not statistically significant.

- **S-1 Card Rotation Test:**
  - Training (pre and post): t = 0.99, p = 0.3481; result is not statistically significant.
  - Game (pre and post): t = 1.80, p = 0.11; result is not statistically significant.

In summary, test comparisons indicate that while there was no difference pre/post for the game group in all measures, there was a significant improvement for the training group in all test measures except for S-1. Figure 5.15 shows performance on all tests before and after the training/game.

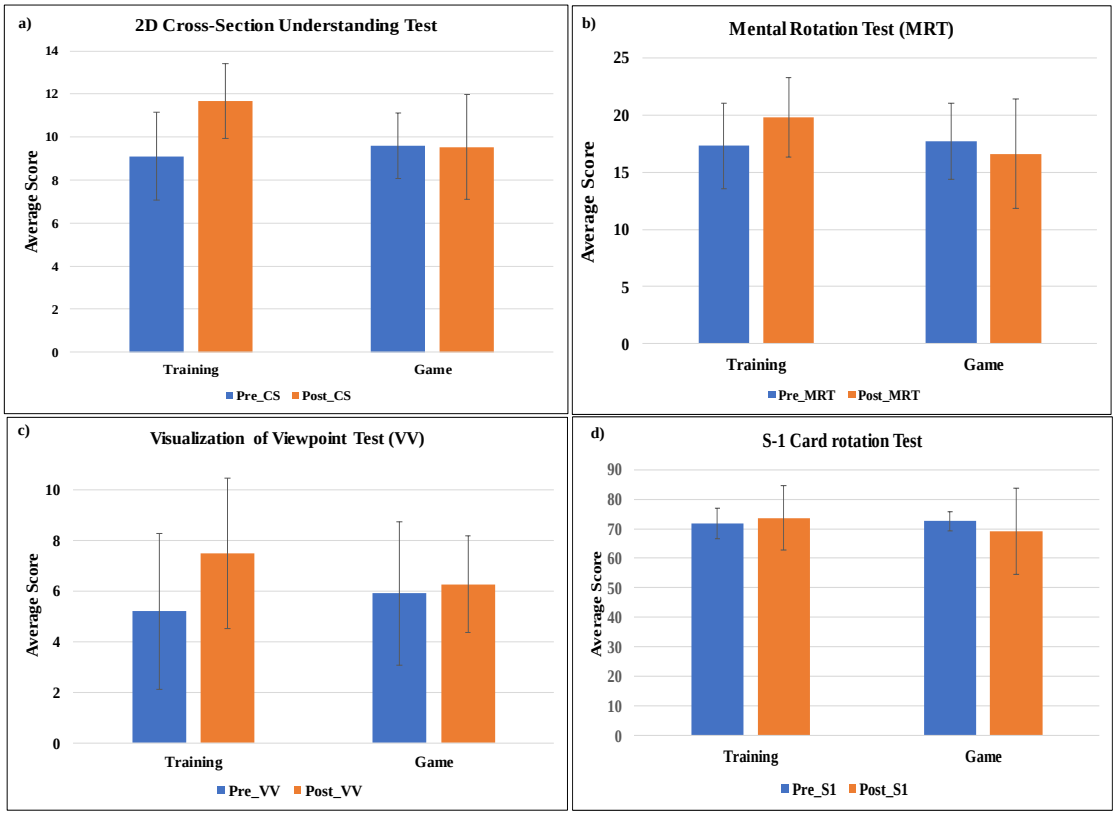


Figure 5.15: Difference in performance on each test measure for each group (training or game). a) 2D cross-section understanding test (max score is 13), there is a significant improvement for the training group but not for the game group; b) Mental Rotation Test (max score is 12), there is a significant improvement for the training group but not for the game group; c) Visualization of Views (max score is 12), there is a significant improvement for the training group but not for the game group; d) Card rotation test S-1 (max score is 80), there is no significant difference on improvement for the training or the game group.

Test Performance Improvements for the Training Group The average test improvement (difference between post-test and pre-test scores) for the training group is relatively higher than the game group (Figure 5.16 a). As shown in Figure 5.18 a) there is also a significant negative correlation between "2D Cross-section" pre-test performance and test performance improvements for the training group ($r = -0.6974$, p-value <0.0001).

Therefore, participants with lower 2D cross-section test scores show more improvements than the participants with higher scores. Based on the "2D cross-section" pre-test scores (the minimum grade for our participants was 5), we grouped the training participants into three subgroups: 1) Low Score (5<score<7); 2) Medium Score (8<score<10); and 3) High Score (11<score<13). As shown in Figure 5.16 b) the highest score improvement was for the "Low Score" group.

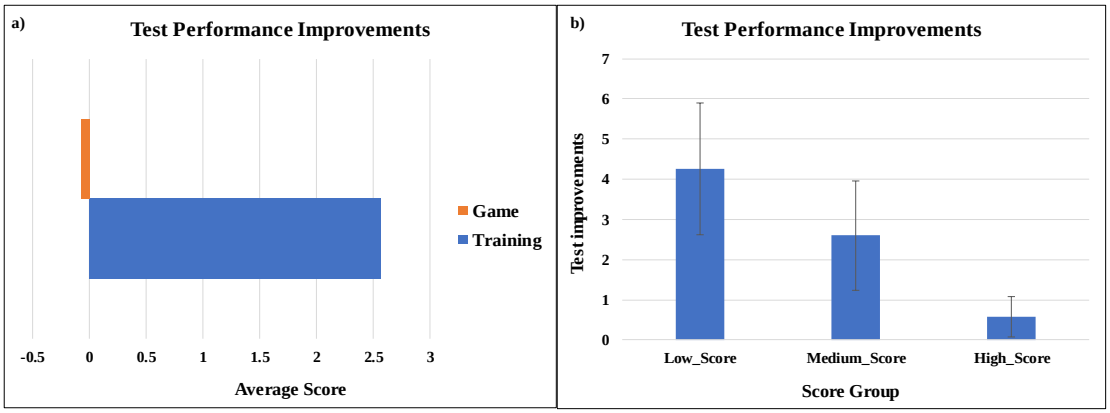


Figure 5.16: Test performance improvements. a) Average score improvements for the training group is significantly higher than the game group (they showed no improvements). b) Performance improvements for the three training subgroups. The Low-score group showed the highest improvements.

5.3.2 Comparing Training and Game Groups Based on Background Questions

The goal of this part is to analyze the participants' answers to the questions of the background questionnaire to gain more insights about their 3D modeling/cross-section experience, and their opinions about the test/ training /game. Figure 5.17 shows the average percentage of ratings for each of the background questions for both the training and game groups.

Experience: Overall there is no significant difference between the training and game group, however on average the game group rated themselves a higher score for "Cross-

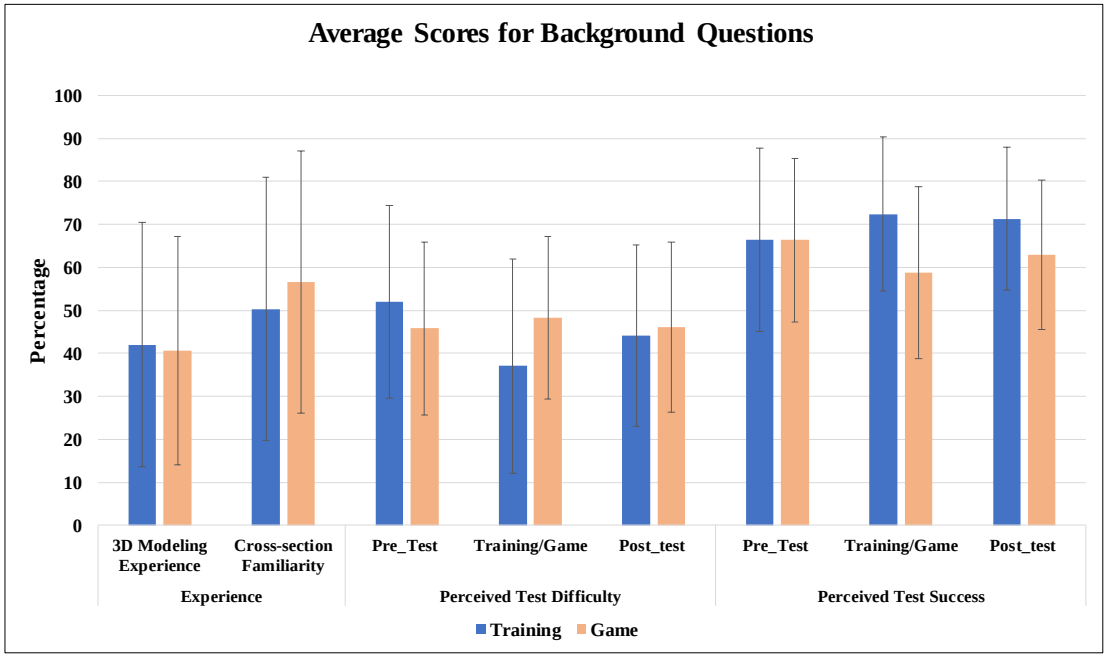


Figure 5.17: Answers to background questions. Training vs. Game group.

section Familiarity". By "Cross-section Familiarity" we mean how familiar someone is with the cross-section concept (before participating in our studies).

A simple one-way ANOVA shows there is a significant effect of 3D modeling experience on test performance for the pre-test: $F(1,57) = 14.47$ $p = 0.00035$.

There is a significant positive correlation between "2D cross-section pre-test performance and "Cross-section Familiarity" ($r = 0.3335$, $p = 0.0098$). There is a significant negative correlation between "2D cross-section test" improvements (difference of post-test and pre-test scores) and "Cross-section Familiarity" for the training group ($r = -0.697$, $p<0.0001$). See Figure 5.18 b,c.

Perceived Test Difficulty: Three is no significant difference between the two test groups, but on average, participants of the training group gave a higher score for the pre-tests difficulty (they found the pre-test to be more difficult), and a lower score for the posts-tests. In addition, participants of the game group, rated the game to be more difficult (comparing to the training tasks).

Perceived Test Success and Training Benefit: On average, training participants rated themselves to be more successful in both the training task and the post-tests. The average perceived benefit of the test on enhancing the cross-section understanding (just for the training group) was 76%. There is a positive correlation between "2D cross-section test" improvements (difference of post-test and pre-test scores) and "Perceived Benefit" of training (r = 0.2395) but the result is not significant. See figure 5.18 d.

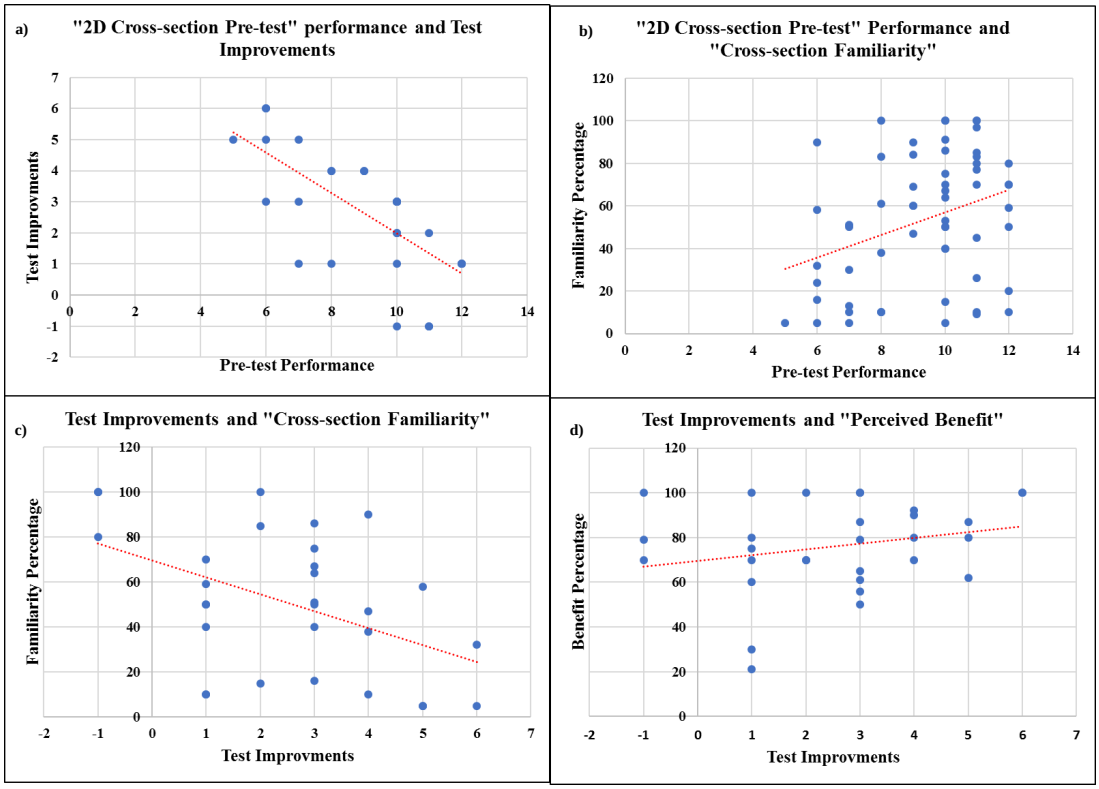


Figure 5.18: Correlations between: a) 2D cross-section pre-test performance and test improvement for the training group (significantly negative); b) 2D cross-section pre-test performance and cross-section familiarity (significantly positive); c) "2D cross-section test" improvement and cross-section familiarity for the training group (significantly negative); d) Non-significant positive correlation between "2D cross-section test" improvements (difference of post-test and pre-test scores) and "Perceived Benefit" for the training group.

5.3.3 Training Tool Features and Test Performance

During the training session, participants worked with the interactive training tool and benefited from its features including: "ask for help", "move/rotate plane", "change view to up/down", "change view to left/right", "ask to show answer", and " check the cross-section". We logged all these interactions of participants with the training tool (e.g., clicking buttons, working with sliders) and tracked their progress throughout the training. Upon completing a training task and providing an answer, we logged the score of participants as well. Table 5.6 shows the average training tasks scores along with the average frequency of interactions with each of the features of the training tool (we call it training tool features usage).

Table 5.6: Training tasks average score and average frequency of interactions with each of the features of the training tool .

	Training Score	**Help**	**Move Plane**	**Rotate Plane**	**Change View Left/right**	**Change View Up/down**	**Show Answer**	**Check Cross-section**
Mean	449.33	15.66	746.4	740.06	30.96	35.9	4.76	24.63
SD	54.88	10.07	311.74	386.32	14.76	13.80	1.49	17.15

We want to know if there are any correlations between the training tool features usage and training task/ test scores. Table 5.7 summarizes these correlations. Except for the "Request to show answer", all the other relationships are not significant. There is a negative significant correlation between the number of requests to see the task correct answers and pre-test scores. There is a positive correlation between request to see the answer and test score improvements. It means participants who more often requested to see the correct answer had lower scores for the 2D cross-section pre-test. After the training they showed more improvements in their test performance. As mentioned, all of the other correlations are not significant. For example, there is a weak negative correlation between the number of "Help" requests and both pre-test and test improvements scores. This shows that those who requested more/less help did not necessarily do better/worse in their test performances. In addition, requests for more help did not change the score improvements.

Simple one-way ANOVA results confirms there is no significant effect of 3D modeling experience on training tool task scores ($F(1,28) = 2.923$ and $p = 0.0984$). There is also no interaction between 3D modeling experience and frequency of use of any of the tool

Table 5.7: Training tasks average score and average frequency of interactions with each of the features of the training tool.

Feature/score	2D Cross-section Pre-test	Test Improvement
Training Tool Task Score	r=0.1525, p>0.05	r=0.068, p>0.05
# Requests for help	r=-0.1347, p>0.05	r=-0.2245, p>0.05
# Moving the plane	r=0.1525, p>0.05	r=-0.2540, p>0.05
# Rotating the Plane	r=0.1525, p>0.05	r=-0.1667, p>0.05
# Changing view to Left/Right	r=0.1525, p>0.05	r=-0.2320, p>0.05
# Changing view to Up/Down	r=0.1525, p>0.05	r=-0.1228, p>0.05
# Requests to show answer	**r= -0.4288, p=0.018**	**r= 0.4222, p=0.010**
# Check Cross-section	r=0.1525, p>0.05	r=-0.2935, p>0.05

features.

5.3.4 Gender Differences Analysis

As mentioned previously, we had 25 females (14 for the training, 11 for the game) and 33 male (16 for training and 17 for the game) participants. 2 participants did not declare their gender.

Based on simple ANOVA results, there is no significant effect of gender difference on overall pre-test performance ($F(1,58) = 2.78$, $p = 0.1$), or test score improvement for the training group ($F(1,28) = 0.332$, $p = 0.569$). However, for each of the test measures, we observed different patterns of average score for males and females. See Figure 5.19.

Gender Differences on Answering the Background Questionnaire: Figure 5.20 and 5.21 show the answers for background questions for male and female participants (game and training). A simple ANOVA indicates that there is no significant difference between male and females except for the perceived benefit of the training tool ($F(1,28) = 4.207$, $p = 0.0497$). On average, for both the training and game group, males rated themselves to be more successful in the tests/training/game. They also rated the tests/training/game to be easier. Females of the game group rated themselves to have more experience working with 3D models and 2D cross-sections.However, they did not rate themselves to be more successful in the tests/training/game. they also found the test to be more difficult. As shown in Figure 5.19 on overage (but not significantly) they outperformed males on all tests expect for the S-1.

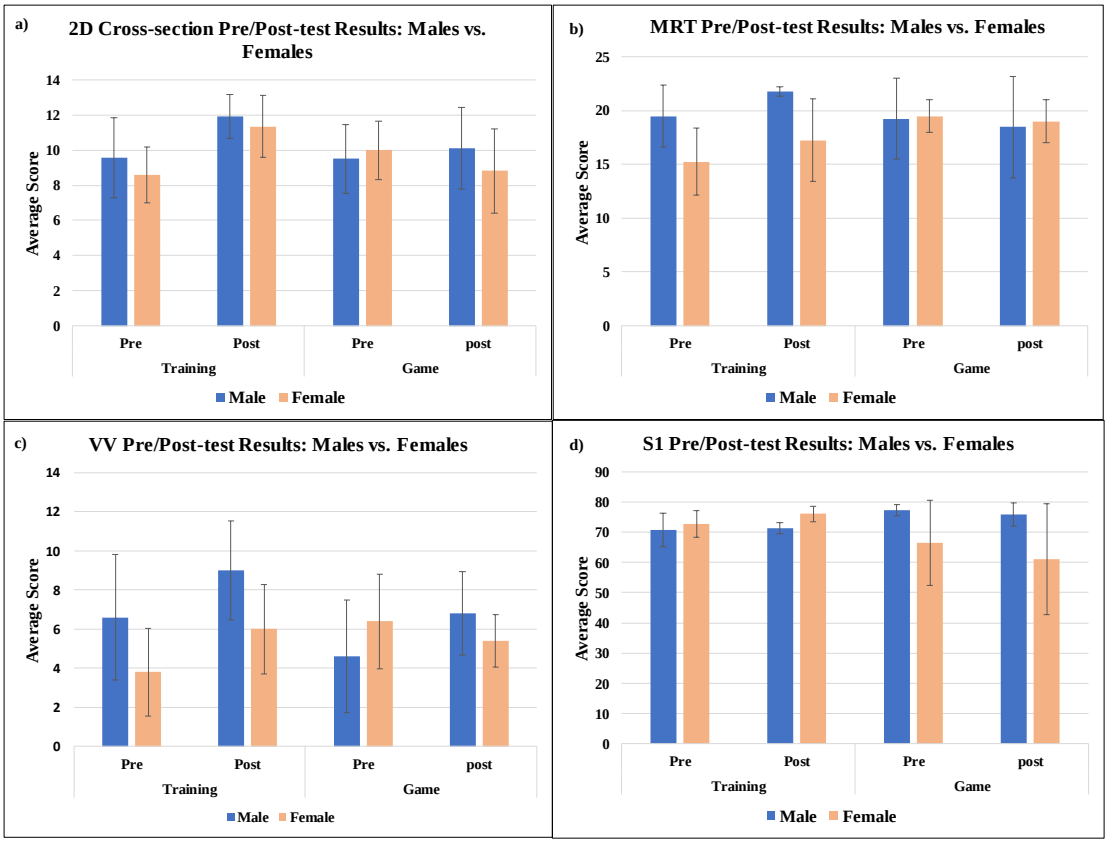


Figure 5.19: Different patterns of score for each test measure based on gender: a) No significant difference between male and females performance for the 2D cross-section test, but males have higher average score for the training, for the game group, initially females outperformed males, but they did not get higher average score for the post-test (results are not significant). b) MRT Test: On average males had better performance in the training group; for the game group women were slightly better. c) For the VV test, males of the training group performed better. While females of the game group were outperformed males in the pre-test, they got lower scores in the post-test. d) S1-Test: for the training females were slightly better, but for the game males outperformed females.

Gender Differences on Using the Training Tool Features: A simple ANOVA indicates that there is a significant effect of gender difference on training tool task scores ($F(1,28) = 9.328$, $p = 0.00491$) . Males significantly outperformed females in completing the training tool tasks. ANOVA results also reveal that there is no difference between

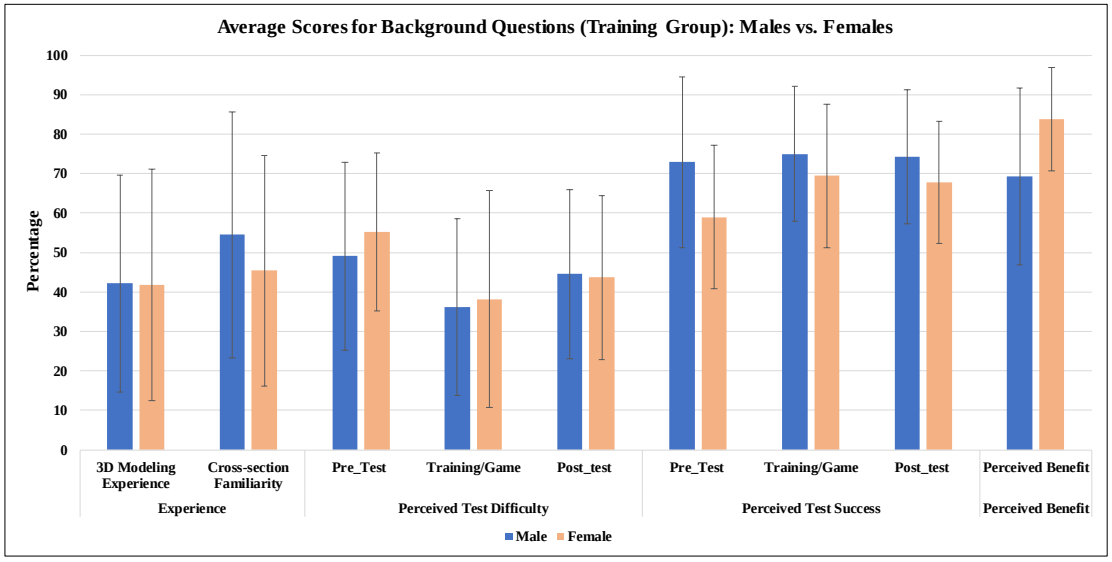


Figure 5.20: Answers to background questions in the training group (males versus females). There is no significant difference between male and females except for the perceived benefit of the training tool. Females found the training tool to be more beneficial for them. In general, males rated themselves to be more successful in the test/training. They also rated the test to be easier.

males and females in using different tool features ($p > 0.21$) except for the frequency of asking to see the correct answer (females requested to see the training task answers more than males and this interaction is slightly significant: $F(1,28) = 4.402$, $p = 0.045$).

Figure 5.22 shows the frequency usage for each of the features in the training tool. On average (but not significantly), female participants have higher frequency of usage of "asking for help", "changing view to left and right", "asking to see the correct answer", and "checking the cross-section". Males more frequently changed the view to up and down and moved/rotated the plane.

5.4 Discussion

In our studies we investigated the effect of using a novel interactive tool to train inferring 2D cross-sections of 3D structures.

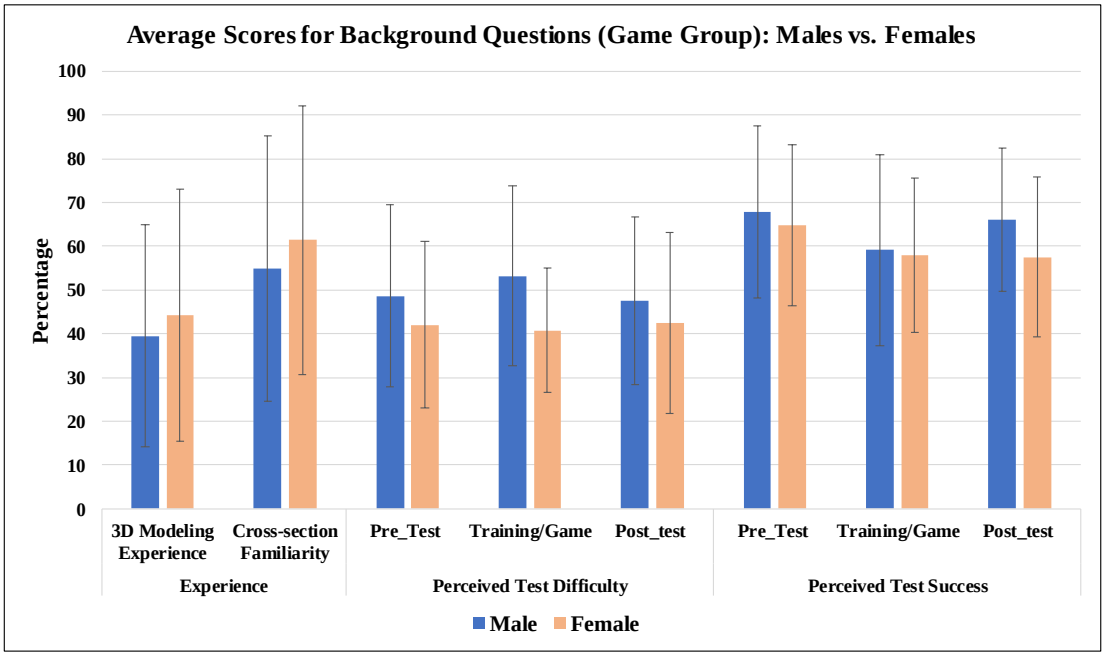


Figure 5.21: Answers to the background questions in the game group (males versus females). There is no significant difference between male and females. Females rated themselves to be more familiar with 2D cross-sections. However, similar to the training group, on average males rated themselves to be more successful in the test/game.

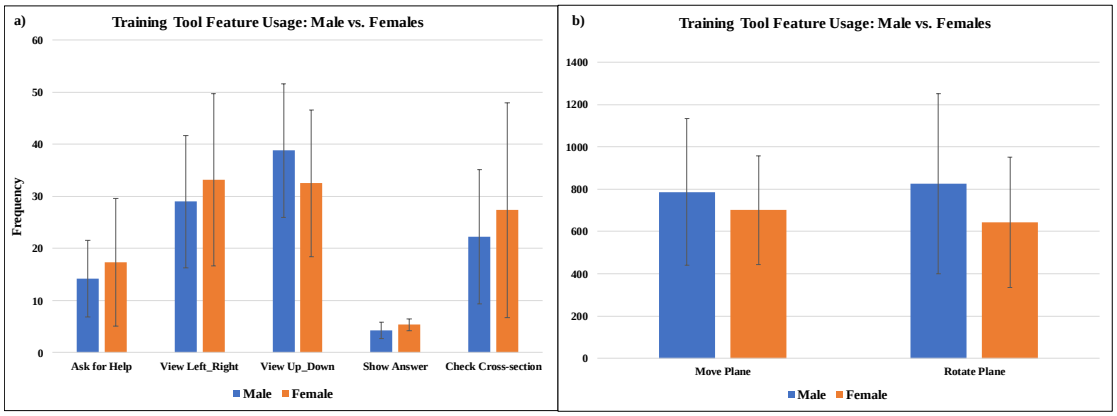


Figure 5.22: Frequency usage for each of the features in the training tool (males versus females.)

[RQ1] Results showed significant performance gains on inferring 2D cross-section for participants of the training group. We did not observe any significant pre/post-test score differences for the game group. Therefore, we can conclude that the tool does effectively help participants to infer 2D cross-sections.

Participants with more 3D modeling/2D cross-section experience showed better performance on both pre and post-tests. In addition, there was a significant negative correlation between pre-test performance and test improvements, meaning that the training was significantly beneficial for those who obtained lower pre-test scores. We detect performance improvements for participants who got high scores in the pre-test. We can conclude that the test was beneficial for everyone, specifically for individuals with lower spatial skills or less 3D modeling experience/cross-section familiarity.

[RQ2] For the training group, we observed a significant performance improvement on the MRT and VV test measures, but not S-1. We also detected that performance on the 2D cross-section test was significantly correlated with the MRT and VV test. This result shows a benefit for working with our training tool and enhancing performance on mental rotations and viewpoint visualization measures. Since our training tool tasks involved multiple mental relations/transition and viewpoint changes, and we provided different tool features for that, we predicted those improvements on MRT and VV test. We are not surprised that there was no training effect in the S-1 test since the primary spatial experience in our training involves mental rotation and viewpoint changing in 3D but not 2D. S-1 test items were all 2D, and the test involve basic 2D transforms.

[RQ3] Average score for the perceived benefit of the training tool was 76%. Also, on average, training group participants gave lower scores for the perceived difficulty of the post-tests, and higher scores for the perceived success. According to all these results we claim that participants found the training to be useful.

[RQ4] Among all of the training tool features, the functionality to "show correct answers" had the most significant impact on performance. Participants who more frequently requested to see the correct answer performed better in the post-test. This result confirms how crucial it is to provide participants with step-by-step correct answers/explanations/feedback for a certain 2D cross-section task. One of our predictions

was to see a positive correlation between number of help requests and test performance improvements. However, we could not detect any significant relationship. One explanation is that the help feature was not that useful for our participants. One possible future direction is therefore to work on the "Help feature" to add more comprehensible instructions and help options.

[RQ5] While previous research results (see Chapter 3) show that males significantly outperformed females in 2D cross-section understanding tasks, in the current study, there was no significant effect of gender difference on overall test performances. A few justifications for this observation are: 1) most of our female participants were graduate students in engineering field and many of them indicated that they do have some 3D modeling experience. 2) For each test measure, we had the limited number of female and male participants. On average (but not significantly) females of the training group did worse on pre-tests (comparing to males), but both females and males showed improvements after the training. So, the training was useful for both genders. On the other hand, while females of the game group ranked themselves at a higher level of experience comparing to the males, and outperformed them in the pre-tests, they did not mange to show improvements or outperform males in the post-tests. One explanation is that they became more tired after playing a challenging game and without any training their spatial skills were not enhanced. Females gave higher score for the perceived benefit of the training and found it to be more useful for them. One interesting observation is that even females with more 3D modeling experience gave lower score for their success in the tests. It seems male participants have more self confidence than females and even found the training to be less useful for themselves.

5.5 Conclusion

Understanding 3D structures through 2D cross-sections is a spatial task that appears both in 3D volume segmentation and many other scientific fields. Similar to other shape understanding and spatial tasks — such as paper folding — performance on this task depends on different spatial skills which varies across the population, and can be improved through training and practice.

We are — long term — interested in creating training tools for 3D volume segmen-

tation. To put our 2D cross-section strategy into practice, in this chapter we introduced our interactive training tool for 2D cross-section understating. We also stepped towards designing and implementing a domain-agnostic training tool.

The results of our experiments showed that the training tool was effective not only in enhancing 2D cross-section understanding, but also improving two other spatial skills: mental rotation and viewpoint visualization. In addition, it produced a verifiable effect across those skills, utilizing only a small window of training. While most research studies on this area implement hours of training, similar to [82] findings, our results also suggest that smaller training windows do have some utility, and future research could possibly explore the actual effectiveness of training durations.

Future Work While our study results proved that the training strategy was effective for engineering students (most of our participants were undergraduate and graduate student in engineering fields), it can also be adapted to the training of simpler spatial skills (e.g., for children) or more complex spatial skills with adults who might most benefit from an increase in spatial skills. One possible future work is to train more participants with diverse backgrounds (gender, age, field of study, level of education), evaluate their performance, and obtain more generalizable results.

The next step for this research is to use a customized version of the training tool in medical/research labs and evaluate its effectiveness in ongoing manual 3D volume segmentation tasks.

Chapter 6: Conclusion and Future Work

In this research, we investigated 3D volume segmentation as a human-computer interaction paradigm to understand and classify human factors involved in the process. Existing 3D segmentation design and evaluation methods pay little attention to the role of humans, their mental models and how they perform low-level tasks and define the higher level criteria while doing 3D volume segmentation. Our research represents a major departure from existing research on either segmentation algorithms or tools.

We proposed employing formative studies to build an understanding of the segmentation process in the context of expert users performing real segmentation tasks. For that, we developed a hybrid protocol that blends elements from existing qualitative and quantitative human-computer interaction methods so that it can be utilized in the field studies (Chapter 2). To ensure reliable study results, we focused on rigorous data collection from several sources (video, audio, eye-gaze) and cross-validation between multiple data streams. Initially, our focus was not on statistical sample analysis but rather on a repeatable coding scheme to identify particular patterns both within and between participants. Our subsequent analysis focused on capturing both low-level (micro) actions and higher-level (macro) tasks. We demonstrated that our micro-task coding scheme effectively captures –in a quantitative manner– the strategies, tasks, and behaviors observed qualitatively. Also, using macro-task classifications, we successfully identified and visualized macro-task flow patterns (Chapter 3).

As we conducted field studies and gathered rich data from observation sessions, we could only focus on small sample size (10 participants). This may limit external validity, and our results may not be generalized to different populations and situations. However, our participants are selected in a way to cover a range of tasks, data sets, tools, and expertise to simulate real world as much as possible. Regarding construct validity, we used multiple sources of data, including eye-tracking, observation video/audio, and field notes. Our study protocol and all the operations of the field studies are well-documented, which makes the studies repeatable, ensuring reliability.

In-depth formative study results also confirmed that understanding 3D shapes through

cross-sections is a necessary mental task that appears in 3D volume segmentation. Similar to other spatial tasks (such as paper folding), performance on 2D cross-section understating varies across the population, and can be improved through training and practice.

Being interested in creating training tools for 3D volume segmentation, we initially modified (and evaluated) an existing cross-section performance measure in the context of our intended application. Our primary adaptations were 1) to use 3D stimuli (instead of 2D) to more accurately capture the real-world application and 2) evaluate performance on 3D biological shapes relative to the 3D geometric shapes. Study results showed that our 3D stimuli was effective (Chapter 4).

We then developed a novel classification for defining a range of difficulty in 2D cross-section understanding tasks. Using this range of difficulty, we created our modified 2D cross-section test instrument, with question items of only organic structures that capture skill sets (Chapter 4).

Finally, to put our 2D cross-section strategy into practice, we designed and implemented a domain-agnostic training tool. We conducted users studies and used our 2D cross-section test instrument to evaluate the training tool. Study results showed that the training was effective for both female and male participants (mainly from engineering fields with various level of experience for 3D modeling). Individuals with lower spatial skills or less 3D modeling experience benefited more from the training (Chapter 5).

Future Work In future we can easily customize the training tool (e.g, by adding simpler/more complex tasks, 3D models, tool features) for different training purposes in various fields including science and engineering. A customized version of our training tool will also be available to our research collaborators (segmentation research labs) targeting novice segmenters. The next step for this research is to evaluate the effectiveness of the customized training tool in various ongoing manual 3D volume segmentation tasks.

Bibliography

[1] Medical image processing, analysis and visualization.

[2] Scott T Acton and Nilanjan Ray. Biomedical image analysis: Segmentation. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 4(1):1–108, 2009.

[3] Abdulla Aldandarawy. Simple cross section shaders. `https://unity3d.com/`, 2016.

[4] Richard Arias-Hernández, John Dill, Brian Fisher, and Tera Marie Green. Visual analytics and human-computer interaction. *interactions*, 18(1):51–55, 2011.

[5] Electronic Arts. Word whomp. `https://www.amazon.com/Word-Whomp-Deluxe-PC/dp/B000PE0H9U`, 2011.

[6] AutoDesk. Autodesk 3ds max. `https://www.autodesk.com/products/3ds-max/overview`, 2018.

[7] Maryann Baenninger and Nora Newcombe. The role of experience in spatial test performance: A meta-analysis. *Sex roles*, 20(5-6):327–344, 1989.

[8] Balsamiq. A rapid wireframing tool. `https://balsamiq.com/`, 2018.

[9] Irving Biederman and Margaret M Shiffrar. Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):640, 1987.

[10] Ann Blandford, Dominic Furniss, and Stephann Makri. Qualitative hci research: Going behind the scenes. *Synthesis Lectures on Human-Centered Informatics*, 9(1):1–115, 2016.

[11] Erwin H Brinkmann. Programed instruction as a technique for improving spatial visualization. *Journal of Applied Psychology*, 50(2):179, 1966.

[12] Andrew Bzostek, G Ionescu, Lionel Carrat, Catherine Barbe, Olivier Chavanon, and Jocelyne Troccaz. Isolating moving anatomy in ultrasound without anatomical knowledge: Application to computer-assisted pericardial punctures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1041–1048. Springer, 1998.

[13] Luigi Franco Cazzaniga, Maria Antonella Marinoni, Alberto Bossi, Ernestina Bianchi, Emanuela Cagna, Dorian Cosentino, Luciano Scandolaro, Marica Valli, and Milena Frigerio. Interphysician variability in defining the planning target volume in the irradiation of prostate and seminal vesicles. *Radiotherapy and oncology*, 47(3):293–296, 1998.

[14] Julia H Chariker, Farah Naaz, and John R Pani. Computer-based learning of neuroanatomy: A longitudinal study of learning, transfer, and retention. *Journal of educational psychology*, 103(1):19, 2011.

[15] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3d mesh segmentation. In *ACM Transactions on Graphics (TOG)*, volume 28, page 73. ACM, 2009.

[16] Isabelle D Cherney. Mom, let me play more computer games: They improve my mental rotation skills. *Sex Roles*, 59(11-12):776–786, 2008.

[17] Richard E. Clark. Design document for a guided experiential learning course. Final report on contract DAAD 19-99-D-0046-0004 from TRADOC to the Institute for Creative Technology and the Rossier School of Education., 2004.

[18] Richard E. Clark. Training aid for cognitive task analysis. Technical report produced under contract ICT 53-0821-0137W911NF-04-D-0005 from the Institute for Creative Technologies to the Center for Cognitive Technology, University of Southern California., 2006.

[19] Richard E. Clark, D. Feldon, Jeroen JG van Merrinboer, Kenneth Yates, and Sean Early. Cognitive task analysis. *Handbook of research on educational communications and technology*, 3:577–593, 2008.

[20] Cheryl A Cohen and Mary Hegarty. Sources of difficulty in imagining cross sections of 3d objects. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, pages 179–184. Cognitive Science Society Austin TX, 2007.

[21] Cheryl A Cohen and Mary Hegarty. Inferring cross sections of 3d objects: A new spatial thinking test. *Learning and Individual Differences*, 22(6):868–874, 2012.

[22] Cheryl A Cohen and Mary Hegarty. Visualizing cross sections: Training spatial thinking using interactive animations and virtual objects. *Learning and Individual Differences*, 33:63–71, 2014.

[23] National Research Council, Geographical Sciences Committee, et al. *Learning to think spatially*. National Academies Press, 2005.

[24] Richard T Duesbury et al. Effect of type of practice in a computer-aided design environment in visualizing three-dimensional objects from two-dimensional orthographic projections. *Journal of Applied Psychology*, 81(3):249, 1996.

[25] Jan Egger, Tina Kapur, Andriy Fedorov, Steve Pieper, James V. Miller, Harini Veeraraghavan, Bernd Freisleben, Alexandra J. Golby, Christopher Nimsky, and Ron Kikinis. GBM volumetry using the 3d slicer medical image computing platform. *Scientific Reports*, 3, 2013.

[26] Zhiw ei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. The validity of the stimulated retrospective think-aloud method as measured by eye tracking, 2006.

[27] John Eliot and Ian Macfarlane Smith. *An international directory of spatial tests*. Cengage Learning Emea, 1983.

[28] Sanne Elling, Leo Lentz, and Menno de Jong. Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1161–1170. ACM, 2011.

[29] Jing Feng, Ian Spence, and Jay Pratt. Playing an action video game reduces gender differences in spatial cognition. *Psychological science*, 18(10):850–855, 2007.

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

[31] Franca Foppiano, Claudio Fiorino, Giovanni Frezza, Carlo Greco, Riccardo Valdagni, AIRO National Working Group on Prostate Radiotherapy, et al. The impact of contouring uncertainty on rectal 3d dose–volume data: Results of a dummy run in a multicenter trial (airopros01–02). *International Journal of Radiation Oncology* Biology* Physics*, 57(2):573–579, 2003.

[32] Jordi Freixenet, Xavier Muñoz, David Raba, Joan Martí, and Xavier Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *European Conference on Computer Vision*, pages 408–422. Springer, 2002.

[33] John W French, Leighton A Price, and Ruth B Ekstrom. *Manual for kit of reference tests for cognitive factors*. Educational Testing Service, 1963.

[34] Kristin M Gagnier, Kinnari Atit, Carol J Ormand, and Thomas F Shipley. Comprehending 3d diagrams: Sketching to support spatial reasoning. *Topics in cognitive science*, 9(4):883–901, 2017.

[35] Andreas Gegenfurtner, Anna Siewiorek, Erno Lehtinen, and Roger Saljo. Assessing the quality of expertise differences in the comprehension of medical visualizations. *Vocations and Learning*, 6(1):37–54, 2013.

[36] Helena BP Gerson, Sheryl A Sorby, Anne Wysocki, and Beverly J Baartmans. The development and assessment of multimedia software for improving 3-d spatial visualization skills. *Computer Applications in Engineering Education*, 9(2):105–113, 2001.

[37] Cindy Grimm and Pushkar Joshi. Just drawit: a 3d sketching system. In *Proceedings of the International Symposium on Sketch-Based Interfaces and Modeling*, pages 121–130, 2012.

[38] Gillian R Hayes. The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(3):15, 2011.

[39] Mary Hegarty. Components of spatial intelligence. In *Psychology of Learning and Motivation*, volume 52, pages 265–297. Elsevier, 2010.

[40] Mary Hegarty, Madeleine Keehner, Cheryl Cohen, Daniel R Montello, and Yvonne Lippa. The role of spatial cognition in medicine: Applications for selecting and training professionals. *Applied spatial cognition*, pages 285–315, 2007.

[41] Mary Hegarty, Madeleine Keehner, Peter Khooshabeh, and Daniel R Montello. How spatial abilities enhance, and are enhanced by, dental education. *Learning and Individual Differences*, 19(1):61–70, 2009.

[42] Mary Hegarty and D Waller. Individual differences in spatial abilities. *The Cambridge handbook of visuospatial thinking*, pages 121–169, 2005.

[43] Michelle Holloway, Anahita Sanandaji, Deniece Yates, Amali Krigger, Ross Sowell, Ruth West, and Cindy Grimm. Guided structure-aligned segmentation of volumetric data. In *International Symposium on Visual Computing*, pages 307–317. Springer, 2015.

[44] Karen Holtzblatt and Sandra Jones. *Contextual inquiry: a participatory technique for system design*, pages 177–210. Lawrence Erlbaum Associates, Hillsdale, 1993.

[45] Lennox Hoyte, Wen Ye, Linda Brubaker, Julia R Fielding, Mark E Lockhart, Marta E Heilbrun, Morton B Brown, and Simon K Warfield. Segmentations of mri images of the female pelvic floor: A study of inter-and intra-reader reliability. *Journal of Magnetic Resonance Imaging*, 33(3):684–691, 2011.

[46] Aulikki Hyrskykari, Saila Ovaska, Päivi Majaranta, Kari-Jouko Räihä, and Merja Lehtinen. Gaze path stimulation in retrospective think-aloud. *Journal of Eye Movement Research*, 2(4):1–18, 2008.

[47] Luis Ibanez, Will Schroeder, Lydia Ng, and Josh Cates. The itk software guide second edition updated for itk version 2.4. *Download: http://www. itk. org*, 2003.

[48] Tao Ju, Qian-Yi Zhou, and Shi-Min Hu. Editing the topology of 3d models by sketching. In *ACM SIGGRAPH 2007 Papers*, SIGGRAPH '07, New York, NY, USA, 2007. ACM.

[49] Yael Kali and Nir Orion. Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 33(4):369–391, 1996.

[50] Madeleine Keehner, Mary Hegarty, Cheryl Cohen, Peter Khooshabeh, and Daniel R Montello. Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive science*, 32(7):1099–1132, 2008.

[51] Madeleine Keehner and Peter Khooshabeh. Computerized representations of 3d structure: How spatial comprehension and patterns of interactivity differ among learners. In *AAAI Spring Symposium: Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance*, pages 12–17, 2005.

[52] Peter Khooshabeh and Mary Hegarty. Inferring cross-sections: When internal visualizations are more important than properties of external visualizations. *Human–Computer Interaction*, 25(2):119–147, 2010.

[53] Gary A Klein, Roberta Calderwood, and Donald Macgregor. Critical decision method for eliciting knowledge. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(3):462–472, 1989.

[54] Elizabeth A. Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010.

[55] U Kuckartz. Maxqda: Qualitative data analysis. *Berlin: VERBI software*, 2007.

[56] Susanne P Lajoie. Individual differences in spatial ability: Developing technologies to increase strategy awareness and skills. *Educational Psychologist*, 38(2):115–125, 2003.

[57] Anthony J Levinson, Bruce Weaver, Sarah Garside, Holly McGinn, and Geoffrey R Norman. Virtual reality and brain anatomy: a randomised trial of e-learning instructional designs. *Medical Education*, 41(5):495–501, 2007.

[58] Rui Li, Jeff Pelz, Pengcheng Shi, Cecilia Ovesdotter Alm, and Anne R. Haake. Learning eye movement patterns for characterization of perceptual expertise. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 393–396, New York, NY, USA, 2012. ACM, ETRA '12.

[59] Marcia C Linn and Anne C Petersen. Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, pages 1479–1498, 1985.

[60] Mark G McGee. Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological bulletin*, 86(5):889, 1979.

[61] Eric N Mortensen and William A Barrett. Interactive segmentation with intelligent scissors. *Graphical models and image processing*, 60(5):349–384, 1998.

[62] Tamara Munzner, Chris Johnson, Robert Moorhead, Hanspeter Pfister, Penny Rheingans, and Terry S Yoo. Nih-nsf visualization research challenges report summary. *IEEE Computer Graphics and Applications*, 26(2):20–24, 2006.

[63] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic Books (AZ), 2013.

[64] Lynn Okagaki and Peter A Frensch. Effects of video game playing on measures of spatial performance: Gender effects in late adolescence. *Journal of applied developmental psychology*, 15(1):33–58, 1994.

[65] S. D. Olabarriaga and A. W. Smeulders. Interaction in the segmentation of medical images: a survey. *Medical Image Analysis*, 5(2):127–142, 2001.

[66] Nir Orion, David Ben-Chaim, and Yael Kali. Relationship between earth-science education and spatial visualization. *Journal of Geoscience Education*, 45(2):129–132, 1997.

[67] John R Pani, John A Jeffres, Gordon T Shippey, and Karen J Schwartz. Imagining projective transformations: Aligned orientations in spatial organization. *Cognitive Psychology*, 31(2):125–167, 1996.

[68] Michael Peters, Bruno Laeng, Kerry Latham, Marla Jackson, Raghad Zaiyouna, and Chris Richardson. A redrawn vandenberg and kuse mental rotations test-different versions and factors that affect performance. *Brain and cognition*, 28(1):39–58, 1995.

[69] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation 1. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

[70] Marios Pittalis and Constantinos Christou. Types of reasoning in 3d geometry thinking and their relation with spatial ability. *Educational Studies in Mathematics*, 75(2):191–212, 2010.

[71] Pixologic. Zbrush. `http://pixologic.com/`, 2018.

[72] Miranda Poon, Ghassan Hamarneh, and Rafeef Abugharbieh. Efficient interactive 3d livewire segmentation of complex objects with arbitrary topology. *Computerized Medical Imaging and Graphics*, 32(8):639–650, 2008.

[73] Claudia Quaiser-Pohl. The mental cutting test" schnitte" and the picture rotation test-two new measures to assess spatial ability. *International Journal of Testing*, 3(3):219–231, 2003.

[74] Qualtric. Qualtrics: an industry leading web-based survey system. `http://main.oregonstate.edu/qualtrics`, 2018.

[75] Anjana Ramkumar, Jose Dolz, Hortense A Kirisli, Sonja Adebahr, Tanja Schimek-Jasch, Ursula Nestle, Laurent Massoptier, Edit Varga, Pieter Jan Stappers, Wiro J Niessen, et al. User interaction in semi-automatic segmentation of organs at risk: A case study in radiotherapy. *Journal of digital imaging*, 29(2):264–277, 2016.

[76] K Rochford. Spatial learning disabilities and underachievement among university anatomy students. *Medical education*, 19(1):13–26, 1985.

[77] Yvonne Rogers. New theoretical approaches for hci. *Annual review of information science and technology*, 38(1):87–143, 2004.

[78] Jean Russell-Gebbett. Skills and strategies–pupils' approaches to three-dimensional problems in biology. *Journal of Biological Education*, 19(4):293–298, 1985.

[79] Anahita Sanandaji, Cindy Grimm, and Ruth West. How experts mental model affects 3d image segmentation. In *Proceedings of the ACM Symposium on Applied Perception*, pages 135–135. ACM, 2016.

[80] Anahita Sanandaji, Cindy Grimm, and Ruth West. Inferring cross-sections of 3d objects: a 3d spatial ability test instrument for 3d volume segmentation. In *Proceedings of the ACM Symposium on Applied Perception*, page 13. ACM, 2017.

[81] Anahita Sanandaji, Cindy Grimm, Ruth West, and Max Parola. Where do experts look while doing 3d image segmentation. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 171–174. ACM, 2016.

[82] Christopher A Sanchez. Enhancing visuospatial performance through video game training to increase learning in visuospatial science domains. *Psychonomic Bulletin & Review*, 19(1):58–65, 2012.

[83] Will Schroeder, Ken Martin, and Bill Lorensen. An object-oriented approach to 3-d graphics, 2005.

[84] Katherine Schultz. The contribution of solution strategy to spatial performance. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(4):474, 1991.

[85] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.

[86] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010.

[87] Leen-Kiat Soh and Costas Tsatsoulis. Learning methodologies and discriminating visual cues for unsupervised image segmentation. In *Seventeenth International Conference on Machine Learning: Workshop on Machine Learning of Spatial Knowledge*, Palo Alto, CA, 2000.

[88] Sheryl A Sorby. Developing 3-d spatial visualization skills. *Engineering Design Graphics Journal*, 63(2), 2009.

[89] R. Sowell, L. Liu, T. Ju, C. Grimm, C. Abraham, G. Gokhroo, and D. Low. Volume viewer: An interactive tool for fitting surfaces to volume data. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling*, SBIM '09, pages 141–148. ACM, 2009.

[90] Ross Taylor Sowell. Modeling surfaces from volume data using nonparallel contours. 2012.

[91] Detlev Stalling, Malte Westerhoff, and Hans-Christian Hege. 38–amira: a highly interactive system for visual data analysis. *Visualization Handbook*, 2005.

[92] Shanhui Sun. Automated and interactive approaches for optimal surface finding based segmentation of medical image data. 2012.

[93] Lindsay Anne Tartre. Spatial orientation skill and mathematical problem solving. *Journal for Research in Mathematics Education*, pages 216–229, 1990.

[94] Melissa S Terlecki, Nora S Newcombe, and Michelle Little. Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied cognitive psychology*, 22(7):996–1013, 2008.

[95] Jayaram K. Udupa, Vicki R. LeBlanc, Ying Zhuge, Celina Imielinska, Hilary Schmidt, Leanne M. Currie, Bruce E. Hirsch, and James Woodburn. A framework for evaluating image segmentation algorithms. 30(2):75–87, 2006.

[96] Unity. The ultimate game development platform. `https://unity3d.com/`, 2018.

[97] David H Uttal, Nathaniel G Meadow, Elizabeth Tipton, Linda L Hand, Alison R Alden, Christopher Warren, and Nora S Newcombe. The malleability of spatial skills: A meta-analysis of training studies. *Psychological bulletin*, 139(2):352, 2013.

[98] Jan Van de Steene, Nadine Linthout, Johan de Mey, Vincent Vinh-Hung, Cornelia Claassens, Marc Noppen, Arjan Bel, and Guy Storme. Definition of gross tumor volume in lung cancer: inter-observer variability. *Radiotherapy and oncology*, 62(1):37–49, 2002.

[99] Jarke J van Wijk. Views on visualization. *IEEE transactions on visualization and computer graphics*, 12(4):421–432, 2006.

[100] Steven G Vandenberg and Allan R Kuse. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, 47(2):599–604, 1978.

[101] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[102] Ruth West, Cindy Grimm, Max Parola, Meghan Kajihara, Anahita Sanandaji, Kathryn Hays, Luke Hillard, and Brandon Lane. Eliciting tacit expertise in 3d volume segmentation. In *The 9th International Symposium on Visual Information and Communication and Interaction (VINCI 2016)*. ACM, 2016 (to appear).

[103] Oliver Wirjadi. *Survey of 3d image segmentation methods*, volume 35. ITWM, 2007.

[104] Robert K Yin. *Case study research: Design and methods*. Sage publications, 2013.

[105] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128, 2007.

[106] Hui Zhang, Jason E Fritts, and Sally A Goldman. Image segmentation evaluation: A survey of unsupervised methods. *computer vision and image understanding*, 110(2):260–280, 2008.

[107] S Zimeras. *Segmentation Techniques of Anatomical Structures with Application in Radiotherapy Treatment Planning*. INTECH Open Access Publisher, 2012.

[108] Ming Zou, Michelle Holloway, Nathan Carr, and Tao Ju. Topology-constrained surface reconstruction from cross-sections. *ACM Transactions on Graphics (TOG)*, 34(4):128, 2015.

APPENDICES

Appendix A: Study Materials

- The link to the supplementary document for "Segmentation Study Materials" is: `https://1drv.ms/b/s!An-DJx9b6pzP33IMl9a9JPhq6CUK`
- The link to the "2D Cross-section Understating Test Version 1" is: `https://1drv.ms/b/s!An-DJx9b6pzP33YHdWArXwgndfiZ`
- The link to the "2D Cross-section Understating Test Version 2, Part 1" is: `https://1drv.ms/b/s!An-DJx9b6pzP33N-k-2KAXz8F_hV`
- The link to the "2D Cross-section Understating Test Version 2, Part 2" is: `https://1drv.ms/b/s!An-DJx9b6pzP33VJZARL2jvOQSpv`
- The link to the "Training Tool Tutorial" document is: `https://1drv.ms/b/s!An-DJx9b6pzP33TK2KNci-KbdOct`