ABSTRACT

Title of dissertation: SIMULATING REALITY:
TRAINING CITIZEN SCIENTISTS
TO JUDGE STREAM HABITATS
IN MULTISENSORY VIRTUAL REALITY

Alina Goldman Striner
Doctor of Philosophy, 2019

Dissertation directed by: Jennifer Preece
School of Information Studies

Citizen science is a form of crowdsourcing that allows volunteers to participate in scientific data collection and analysis. Many citizen scientists are engaged and motivated by science-based learning and discovery, but high training costs and limited resources often result in volunteers participating in unskilled work, leading to boredom and disengagement.

Advances in immersive virtual reality (VR) have created opportunities to recreate physical environments with minimal cost, making it possible to train citizen scientists to make qualitative experiential judgments usually reserved for domain experts. This research trains citizen scientists to assess outdoor stream habitats using StreamBED VR, a multisensory VR training platform.

This research offers the following contributions:

1. A study of how expert and novice water monitors make qualitative assessments of outdoor stream habitats using an EPA qualitative protocol. The research

found that experts *develop intuitive judgments* of quality, *use multisensory environmental information* to make judgments, and *construct past and future narratives* of streams using environmental characteristics.

2. Iterative design of the Ambient Holodeck multisensory system, and a study of how ambient sensory information impacts observation skills. The research found that multisensory information *increased the number of observations* participants made, and *positively affected engagement and immersion.*

3. Iterative design of the StreamBED VR training platform and two studies; the former explores how qualitative assessment skills can be taught in VR, and the latter considers how training in VR, with and without Multisensory cues, compares to a PowerPoint (PPT) baseline. Study results found although VR participants were *more excited to continue training* than PPT participants, *Standard VR and PPT participants scored closest to an expert gold standard, performing significantly better than Multisensory VR* participants.

This research concludes that VR has the potential to train qualitative assessment tasks, but qualifies that training design is multifaceted and complex, full of theoretical learning considerations and practical challenges. Further, VR realism can be a powerful tool for training, but is only effective when training cues clearly parallel assessment tasks.

SIMULATING REALITY: TRAINING CITIZEN SCIENTISTS TO JUDGE STREAM HABITATS IN MULTISENSORY VIRTUAL REALITY

by

Alina Goldman Striner

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Jennifer Preece, Chair/Advisor
Dr. Kari Kraus
Dr. Beth Bonsignore
Dr. Niklas Elmqvist
Dr. Amitabh Varshney, Dean's Representative

Acknowledgments

Thank you to my adviser, Dr. Jenny Preece, and to my dissertation committee, Dr. Kari Kraus, Dr. Beth Bonsignore, Dr. Niklas Elmqvist, and Dr. Amitabh Varshney. I really appreciate your time and feedback throughout this long dissertation process.

Thank you also to the StreamBED VR team, Parv Rustogi, Pranav Kulkarni, Aashrey Sharma, Connor Petrelle, Sharan Hedge, and Zach Rodriguez. Thank you for your development skills and all of your hard work. Likewise, thank you to all of my participants, without whom this work would not have been possible.

Thank you to my wonderful family for their continued support. Ben, I love that you continually challenge me. Mom and Dad, thank you for giving me the opportunity to follow my own path. Grandma and Grandpa, thank you for always believing in me. Carol and Sam, thank you for giving me such a wonderful husband.

Table of Contents

List of Figures

List of Abbreviations

HCI	Human Computer Interaction
HMD	Head Mounted Display
Mulsemedia	Multisensory Media
PPT	PowerPoint
NPC	Non-Player Character
RBP	Rapid Bioassessment Protocol
RTD	Research Through Design
VR	Virtual Reality

Chapter 1: Introduction

Qualitative judgments scaffold the way we speak, listen, interpret, and evaluate the world around us. Unlike procedural judgments, which split into a sequence of followable steps or rules, qualitative judgments do not have discernible or obvious answers, but must be continually weighed within context [47].

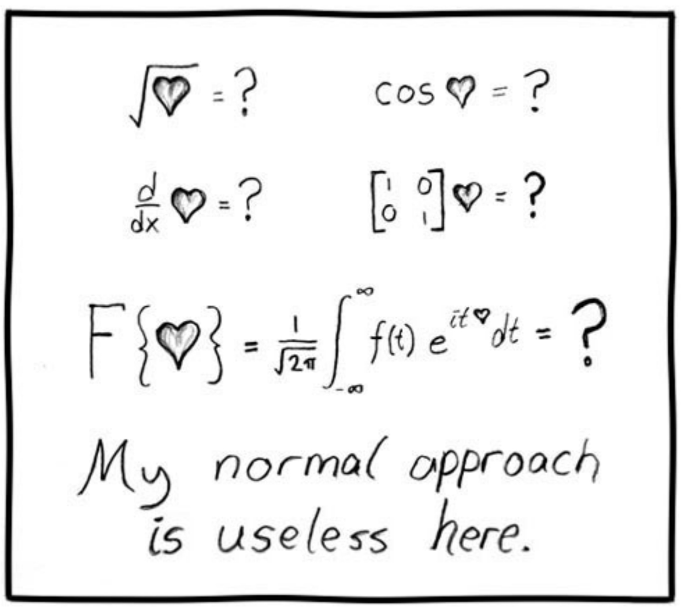

Figure 1.1: Love is a qualitative judgment that is difficult assess through quantitative metrics. Adapted from XKCD [92].

Qualitative judgments are shaped through experiential learning, cyclically fashioned together from multimodal and multisensory information, interpreted, and interwoven into previous experience. For instance, knowing whether someone loves you is a qualitative judgment, shaded by (an often bewildering) combination of glances, embraces, aromas, and intonations (e.g. Figure 1.1). Instead of ambling down a clear, well-trodden path, experiential learners trek through a wilderness of information, wading through data, advice, tips and advertisements, foraging for a blueprint onto which to scaffold their experiences.

1.1 Making Scientific Judgments

Our tendency to notice, examine, and consider events and phenomena is largely guided by our education and rehearsed by our training. In science, a large chasm divides professional scientific observations and reasoning from amateur conjectures. Although scientific fields range and specialize, they are allied by the scientific method's process of systematic observation, measurement, and experimentation to formulate, test, and modify hypotheses.

As effortfull as it is to make everyday qualitative judgments, scientific qualitative judgments are far more laborious and taxing. As with everyday judgments, scientific judgments are the product of accumulated experiences, but are expected to be systematic and reproducible. To make qualitative judgment systematic, training helps researchers scaffold experiences and judgments into a internal framework of meaningful relationships and patterns. Both in everyday and scientific judgments, this framework is never completed, but perpetually patched and refitted with fresh information and experiences [63].

Amateur observers have particular difficulty forming credible judgments of natural environments without domain specific training; novice nature observers are often subject toward ill-fitted or biased interpretations of recognizable phenomena, and may overconfidently appraise their qualifications to make judgments of domains they are superficially familiar with [34].

1.2 Citizen Science

Citizen science is a form of crowdsourcing that involves volunteers in scientific investigation, conservation, education, and activism projects [9]. Citizen science volunteers have varied backgrounds and motivations. In contrast to researchers who spend years working in the field, citizen scientists often have limited experience in the scientific method [18]. Volunteers also have diverse motivations; many have a personal interest in a topic and want to extend their knowledge in the field, some look for acceptance or acknowledgement from the scientific community [18], and others contribute for short-term entertainment [12, 36, 105].

Citizen science has the potential to add value to scientific research, however volunteers often waste project resources; they receive poor training and are tasked with boring, low-skill assignments. Training volunteers for high-level tasks requires a large demand on researcher time and resources, and most citizen science projects are under-resourced and understaffed [123], without means to conduct individual training. Instead, citizen scientists are often given passive training materials [9], and are tasked with unskilled work (e.g. counting, classifying, and transcribing environment features) that requires limited training [123, 18, 105, 36]. While practical, this can lead to boredom and disengagement [38, 65], which may decrease participant retention and data quality [34].

1.3 Training Qualitative Stream Assessment in Virtual Reality

Qualitative stream assessment [6] allows water monitoring groups to visually assess stream habitats to report on the balance between development and environmental protection. This qualitative assessment task is based on a complex ecosystem of related habitat features that are judged in context of one another. Usually reserved for expert monitors, qualitative water monitoring is a high-level assessment task trained by physically experiencing and rating outdoor stream habitats.

This dissertation asks whether qualitative stream assessment skills can be taught to potential citizen scientists through virtual reality (VR). VR affords users with a sense of telepresence, an illusion of actually "being" in a virtual world. Previous research suggests that VR can help learners develop spatial knowledge and help train environmental judgment skills [29]. The goal of my research is to consider whether VR can be employed to train citizen scientists to make qualitative stream assessments the way experts do, giving them practice rating virtual stream habitats.

1.4 Research Questions

My research broadly asks: *How can technology effectively train scientific qualitative judgment tasks?* To undertake this broad question, I consider the following research questions in the domain of citizen science water quality monitoring:

RQ1: *How do expert monitors make qualitative assessments?* This first question endeavours to understand the expert learning and assessment process, considering how expert practices may be transferred to training.

RQ2: *Do multisensory cues help learners make habitat observations?* The second question focuses on the role of multisensory realism in habitat assessment, considering whether ambient sensory information helps learners make better observations of their environment.

RQ3: *How to scaffold learning to support the expert process?* The final question considers how training should be designed to help learners develop the observation and assessment skills of expert monitors.

By addressing these questions, this dissertation contributes new knowledge about the efficacy of training water monitoring through virtual reality.

1.5 Dissertation Structure

The following chapters address and synthesize the research questions through a series of studies and discussion. In Chapter 2, I present relevant theoretical literature, summarizing background research on qualitative judgments, citizen science, virtual reality, and multisensory media technology. Then, in chapter 3, I describe my research methods, first presenting a review of research through design (RTD), then outlining the individual research methods I employed for each study.

Next, I present my four studies. In chapter 4, the *first study* observes and reports on differences in background and training between professional and volunteer monitors, synthesizing findings about volunteer training needs.

In chapter 5, the *second study* designs and evaluates the initial StreamBED VR training that teaches volunteers to make qualitative assessments of stream habitats through physical exploration.

In chapter 6, The *third study* evaluates the role of multisensory realism on observation skills in VR by designing the Ambient Holodeck, that allows users to feel and smell landscape and environmental conditions. The study tests participant observations of stream habitats with and without ambient multisensory cues.

In chapter 7, The *final study* integrates findings from the first three studies, designing the StreamBED VR 2.0 prototype, a collaborative learning experience. The final study tests the effectiveness of this prototype with and without multisensory sensory cues, and against a baseline PowerPoint training.

After describing all four studies, Chapter 8 discusses the value of virtual reality for citizen science training, and appraises the role of multisensory realism and collaborative training on participant learning and engagement. Finally, Chapter 9 discusses applications of my work to citizen science and virtual reality design, considers study limitations, and proposes future research based on my findings.

1.6 Terms

- *Citizen Science*: The practice of including non-scientists in data collection/evaluation.

- *Citizen Scientists:* Volunteers that participate in citizen science projects.
- *Protocol Features:* Components that make up the ratings for the Epifaunal Substrate and Bank Stability scales.
- *Head Mounted Display (HMD)*: A virtual reality display that users wear, such as Oculus Rift.
- *Immersion*: A state of deep engagement during an activity or task.
- *Motivation*: The desire or willingness to complete a task.
- *Multisensory Media or Mulsemedia*: Information communicated through multiple sensory channels, including, visual, auditory, olfactory and haptic inputs.
- *Presence*: A perception of being physically present in a non-physical world.
- *Professional or Expert Monitors*: Researchers that professionally work on water quality monitoring.
- *Research Through Design (RTD):* An HCI research and design method.
- *Rapid Bioassessment Protocol (RBP):* A widely used water monitoring protocol that qualitatively assesses the makeup of stream habitats.
- *Virtual Reality (VR)*: A computer-generated simulation of a three-dimensional image or environment that seems realistic.
- *Water Quality Monitoring*: The practice of monitoring stream habitats in order to assess human impact on watersheds.

Chapter 2: Background

The following chapter presents background research relevant to the dissertation. First, I summarize background literature on *qualitative judgments*, *citizen science*, and *stream monitoring* to contextualize the qualitative judgment task that I explore in this dissertation. Then, I summarize literature on *virtual reality* (VR) and *multisensory media* (mulsemedia) technology.

2.1 Qualitative Judgments

In the following section, I provide an operating definition of qualitative judgments, and provide examples of qualitative judgments in nature. Then, I overview of how they are taught and learned by professionals, and how novices can learn to make such judgments through experience.

2.1.1 Making Qualitative Judgments

Qualitative judgments are intuitive representations of mental models that scaffold the way people speak, listen, and interpret the world around them. In this work, I define making qualitative judgments as using intuitive heuristics [47] to discern relationships between latent variables. Researchers often use these contextual judgments to describe and evaluate knowledge that is difficult to teach and evaluate procedurally [98].

Making qualitative judgments requires evaluators to scaffold knowledge into internal cognitive maps, which chunk interconnected ideas into concepts and models that help them notice, compare, and identify patterns [122]. Over time, this scaffold is shaped by experiential learning [63, 53], cyclically fashioned together from multimodal and multisensory information [52], and from discussion, which is interpreted, and interwoven into previous experience [130].

Making qualitative judgments is difficult and often reserved for professionals. For instance, clinical psychiatrists recommend medical treatments based on qualitative assessments of patient conditions [100], and neurologists use qualitative assessments to characterize patient pain [90]. Similarly, trainers use qualitative methods to evaluate athlete performance [71], and voice teachers use qualitative pedagogical techniques to diagnose vocal problems [86].

Scientists likewise use qualitative heuristics to understand nature. For instance, climate scientists use qualitative "climate proxies," like perennial cherry blossom bloom periods, to assess the cumulative effect of climate change on changes in temperature or rainfall [118]. Likewise, paleoecologists use fossil and sediment proxies to interpret and reconstruct ecosystems and environmental conditions of the past [8]. On a larger scale, astronomers pair qualitative heuristics with quantitative analyses to understand the conditions of unusual cosmic phenomena. For instance, astronomers draw conclusions about supernovae composition using heuristic interpretations of Hubble telescope images of scattered light echoes [108].

2.1.2 Learning to Make Qualitative Judgments

In the natural sciences, learning to make qualitative judgments consists of (1) *noticing* and *observing* phenomena, (2) *comparing* observations to expectations, (3) *synthesizing* observations into patterns, (4) *scaffolding* patterns into a cognitive map [34], and (5) *updating* the cognitive map using new information. First, learners notice phenomena, and make observations; this requires them to know when to ask questions, and what question to ask (*e.g.*"what thing am I looking at?"). Learners then synthesize patterns by comparing observations to internal webs of information [3], and recognizing similarities and differences [20]. Finally, learners scaffold observations into an internal cognitive map, and iteratively update the map by making new observations [63]. Learners also update their cognitive map through discussion, which helps them form shared interpretation of meaning through metaphors, which allow them to transfer information between familiar and new domains [20].

Teaching qualitative judgments is often challenging because experiences are subjective and difficult to surface, examine, compare, and explain [20]. In the examples above, professionals learn to make assessments through experiential learning; they make observations and identify patterns by iteratively getting feedback from patients and students. Unlike professionals who learn from feedback loops, researchers learn to make intuitive judgments by studying related quantitative and qualitative data; for instance, in the astronomy example above, quantitative data about how light scatters helps astronomers visually interpret low resolution telescope images.

Professionals are able to make qualitative judgments because they have more experience—however, with enough experience, amateurs, too, are able to learn to make effective qualitative assessments. For instance, amateur bakers often use qualitative assessments to "troubleshoot" finicky recipes [88]. Similarly, successful amateur investors often use intuitive heuristics to "predict" company valuations and stock market changes [110].

2.2 Engaging Citizen Scientists in Substantive Tasks

Citizen science is a form of crowdsourcing that allows volunteers to collaborate with researchers on scientific data collection and analysis [9]. A long-term goal of environmental citizen science research is to engage local communities in "democratic ownership" of environmental projects to provoke large-scale habitat accountability and civic protection efforts [109].

Literature suggests that many citizen scientists are engaged and motivated by science-based learning and discovery [105], but high training costs and limited resources often result in volunteers participating in only unskilled work. While practical, tasking volunteers with unskilled work can lead to complications, including boredom and disengagement [38, 65], which may decrease data quality and participant retention [34]. Citizen science work is often comprised of unskilled tasks such as documenting, counting, identifying, classifying, and transcribing environment features [12, 123, 18, 105, 22, 36]. Instead of providing scientific education or increasing task skills, projects may engage community members through gamification [11] and novelty [58].

Engaging citizen scientists in substantive, high-level tasks has potential to improve data quality and participant retention [119, 124]. This research proposes engaging those citizen scientists motivated by learning goals with qualitative assessments—high-level assessment tasks focused on naturalistic, inductive interpretations of an environment [96, 31]. A traditional barrier to assigning high-skill tasks to volunteers is that developing the required skills is difficult and time consuming. Qualitative assessments are usually made by professionals who learn through iterative personal experiences. Learning to make qualitative assessments requires access to a variety of heterogeneous environments, which is not practical for volunteers.

2.3 Qualitative Judgments in Stream Monitoring

In the following sections, I first overview the importance of qualitative stream monitoring, then describe the EPA's Rapid Bioassessment Protocol metrics.

2.3.1 Overview of Qualitative Stream Monitoring

Stream ecology groups perform qualitative stream monitoring and assessment to report on the balance between development and environmental protection. For instance, the Maryland Department of Natural Resources uses monitoring to benchmark long term stream quality trends, examine impact of land use on habitat quality and biological diversity, and assess cumulative impacts to streams [113]. Ecology groups also use data from monitoring projects to advocate for land use that protects regional watersheds [6]. To effectively protect watersheds, Roth and Davis [113] describe the need for cost-effective solutions that inform the public about stream conditions and reduce sampling program costs. Training citizen scientist monitors to perform stream monitoring judgment tasks may dually benefit water monitoring agencies by creating a free task force, engaging the public in water quality issues and increasing volunteer retention.

I situate my research in water monitoring for four reasons: (1) a standard protocol is used by many monitoring groups [101], (2) assessments are exclusively made by professionals, (3) assessment does not require specialized tools, making it an ideal task for volunteers, and (4) assessment relies on first-hand experiences at stream sites, which immersive technology has the power to recreate.

2.3.2 Evaluating the EPA's Rapid Bioassessment Protocol (RBP)

This work focuses on the EPA's Rapid Bioassessment Protocol (RBP) [6], a qualitative metric of stream conditions that is used nationally. The RBP is part of a larger monitoring process that includes several quantitative assessment measures. These include measuring PH and temperature, and counting fish and benthic macroinvertebrates, small animals that live among stones, sediments, and aquatic plants, whose presence is indicative of stream quality [6]. Although many watershed monitoring protocols exist, I focus on this protocol because it is employed by many water quality monitoring groups [101].

The RBP is made of 13 measures, described below. Each measure is assessed on a 10 or 20 point scale. An example, channel alteration, is shown in Figure 4.4.

1. *Epifaunal substrate* evaluates opportunity for insect colonization and fish cover
2. *Embeddedness* of cobble and boulders in stream sediment appraises the surface area available to fish and macroinvertebrates for shelter and spawning
3. *Pool substrate characterization* evaluates the mixture of stream bottom substrates
4. *Velocity depth combinations* notes diversity of water velocity and depth patterns
5. *Variability of pool environments* characterizes the diversity of stream pools
6. *Sediment deposition* estimates sediment accumulation at the bottom of a stream
7. *Channel flow status* describes the degree to which a channel is filled with water
8. *Channel alteration* estimates the extent to which a stream's shape has been altered
9. *Frequency of riffles and bends*, judges the heterogeneity of a stream's shape
10. *Channel sinuosity* evaluates the curvature of the stream
11. *Bank stability* considers the extent to which banks have eroded
12. *Bank vegetation protection* values the quality of vegetation protecting the stream

13. *Riparian zone width* delineates the vegetative buffer between a stream bank and runoff pollutants

2.4 Virtual Reality

In the following section, I introduce the benefits of virtual reality (VR), including presence, immersion, enjoyment, and learning. Then, I describe design and research trends to employ the affordances of the technology.

2.4.1 Presence, Immersion, and Learning Benefits

Virtual reality is a computer-generated scenario that simulates physical experiences, allowing users to engage with an artificial world. VR affords users with a sense of telepresence, an illusion of actually "being there" in a virtual world [2]. For instance, Bakdash et al. [4] found that large displays, particularly virtual reality, allow learners to fully master a complex virtual environment by transferring spatial knowledge, helping them effectively keep track of and spatially update where objects are in the environment. Relatedly, Krokos [75] found that virtual memory palaces helped learners recall celebrity faces using environmental context, generalizing that memorable VR experiences can enhance recall of large amount of data.

Further, Cummings [23] summarizes that VR presence affects task judgment, learning, and enjoyment. For instance, Ahn [2] found that giving users an embodied experience of cutting down a tree in VR encouraged them to conserve paper. The effect of presence created through virtual reality has also been documented on emotion; Persky [99] found that playing a violent game in an immersive virtual environment led to increased aggressive feelings and behavior, and Ramalho [106] created a VR experience that helped users sense different levels of object ownership.

2.4.2 Designing for VR Affordances

HCI has only begun to explore the needs and opportunities of this technology. A prominent space considers the affordance of interacting with 360° panoramas. For instance Vermeer [15] and JackInHead [61] used VR to give users "*eyes in the back of their head,*" by dynamically projecting and stitching together 360° images, and IRIDiuM [72] proposed a solution for presenting movie quality graphics with free view-point head motion. Similarly, research has considered how virtual 360° interaction may be experienced in the physical world. For instance, Shadowshooter [129] dynamically created 360° experiences without the need of VR by incorporating a projector worn by the user. Another space considered how objects might be mapped between real and virtual space. For example, Murakami [93] used VR to estimate furniture scaling for interior design [91].

Researchers and practitioners are beginning to use VR to train judgment tasks, such as understanding habitats [28], safety measures [77], and firefighting situations [120]. Ecology researchers have also begun designing for the affordance of VR. A conservation project, Many Eyes on the Wild [7, 67] learned about the habits of jaguars in Peru by collecting 360° images using GoPro cameras and multidirectional sound recordings of the environment. The VR experience allowed virtual scientists to experience remote inaccessible places, and led to better quality information and more confidence in expert analyses.

2.5 Multisensory Media

The following section first introduces the psychological benefits of multisensory media (mulsemedia), and describes HCI efforts to design for multiple senses. The section then overviews literature on learning through multiple senses, and research on sensory information in judgment tasks.

2.5.1 Multisensory Modalities

In contrast to audio-visual media, Ghinea [45] defines “mulsemedia” as three or more assimilated senses. Mulsemedia helps users process and interpret information by creating episodic sensory information [45]. In design, Haverkamp [52] suggests that mulsemedia improves information processing by coactivating unimodal sensory channels on multiple levels of consciousness [46].

Literature also suggests that sensory modalities affect one another. For instance, Krishna et. al. [74] describe the presence of multisensory congruence, showing that sensory experience in different modalities (e.g. touch and smell) can impact one another. Fujisaki [42] and Donley [33] found that combining sensory modalities can impact the perception and quality of an experience, such as using ambient thermal stimuli to moderate body regulation [17].

2.5.2 Designing for Multiple Senses

HCI researchers have begun to explore the viability of integrating multisensory technology into design [94]. For instance, Israr [56] enriched storytelling with haptic feedback, and Iwata [57] simulated the feeling of eating. Relatedly, Spence [115] thoroughly reviewed current technology to transfer smell and taste senses online.

Researchers have also endeavored to integrate multiple sensory experiences into VR. Lopes [79] simulated the physical impact of boxing in VR using electrical muscle simulation, Kiltini [66] explored multisensory and sensorimotor feedback in body ownership, and Gerry [44] superimposed VR on top of physical reality, allowing a novice painter to replicate artist movements on a canvas in VR. HCI research has likewise attempted to replicate ambient experiences; Ambioterm [107] and Amphibian [60] simulated environmental conditions in a VR headset, and Martins [81] conceptualized sensory wine tourism.

2.5.3 Learning through Mulsemedia

Literature suggests that mulsemedia positively impacts learning. For example, Shams and Seitz [114] overview the benefit of sensory inputs on information encoding, recognition and retrieval, cross-modal memory transfer, and reinforcement learning. In light of these benefits, the field of HCI has begun to consider the impact of multisensory cues in education. Yannier [128] found that shaking interactions helped children enjoy learning physics principles, Covaci and Ghinea [19] found that olfactory information and feedback in an education game engaged students, and Zou [133] found that integrating olfaction, airflow, and haptics stimuli increased learner enjoyment.

Research has begun conceptualizing the role of multisensory cues on learning in VR. Early work by Psotka [104] and Dihn [32] suggest that multisensory VR can reduce conceptual load, create salient memories, and increase memory, emotion, and presence for environmental information. Further, Dede (1999) [27] found that multisensory information helped students understand complex scientific models through experiential metaphors and analogies, and displaced intuitive misconceptions.

2.5.4 Mulsemedia in Judgment Tasks

Research suggests that sensory information is particularly important in training judgment tasks because people make observations and process information through different sensory channels [94]. In education, HCI researchers have begun to consider the impact of additional sensory information on perception and learning tasks. For instance, Demate [30] found that olfactory cues could influence people's judgments of facial attractiveness, Yannier [128] and Brooks [13] found that haptics helped improve visualization of complex data sets, and Lee [78] found that visual, auditory, and tactile feedback improved virtual race car drivers' performance.

Notably, in environmental training, Dede (2017) [29] found that VR helped make topographic characteristics of watersheds more noticeable, and suggested that sensory information (e.g. water sounds, weather variables, shifts in grass color) could help learners sense pattern changes. Even though literature suggests that sensory fidelity can impact learning [16], training often limits sensory channels to audio and visual inputs, which can inhibit cognitive model formation [34].

This chapter has introduced key literature in this dissertation: the nature of qualitative judgments, citizen science, and stream monitoring, and advances in virtual reality and multisensory media technology. In the next chapter, I outline my dissertation research methods.

Chapter 3: Overview of Method

The following chapter summarizes my study method. First, I present a review of research through design (RTD), then I outline the individual research methods I employed for each study.

3.1 Research Through Design

This section describes research through design theory and application to HCI. First, I describe RTD philosophy, and the method's use in HCI. Then, I describe the RTD process and outcome measures.

3.1.1 Philosophy

In recent years, HCI has shifted from a narrow focus on usability to more broadly consider the human experience [4]. This has expanded the scope of design to consider "wicked" problems, difficult research questions with unclear or conflicting agendas and messy solutions [4, 117, 131]. Such iterative design may have either a philosophical approach, investigating or embodying a philosophical stance in design, or a grounded approach, focusing on a real world problem [131].

3.1.2 Use in HCI

Research through design produces intermediate knowledge that informs practice [4, 80]. RTD can produce guidelines and heuristics in how to design certain goals [4], but also has potential to produce relevant and rigorous theory [4, 43]. RTD can never produce falsifiable results (a single design cannot be proven incorrect) [43], so criteria for evaluation includes: 1) method and rationale for the selecting research and design methods, 2) research process documentation, 3) novelty and effectiveness of integrating existing theory and practice, and 4) situational contribution and extensibility that advances the state of the art in a research community [4]. RTD has been adapted by HCI to address these wicked problems [4, 43], but RTD literature cautions that to be effective, HCI must embrace design practice as complex "*hit-and-miss…craft-work*" [41] with limitless sources of information, requirements, demands, wants, needs, and opportunities [4, 132].

3.1.3 Process and Outcomes

Research through design methods include observation, prototyping, in-depth interviews, contextual inquiry, usability studies, and master-apprentice models [4].

These methods support new knowledge creation through a design cycle of reflection and annotation that help designers ideate, iterate, critique potential solutions, and continually reframe problems [4]. To reflect on problems and reframe solutions, RTD uses prototypes, products, and models to codify understanding of particular situations [131]; iterative design considers what a potential future might look like by creating an "ultimate particular" [117], an object that solves a particular design need [43]. RTD may also include materials research and development, pushing the boundaries of what a material can achieve, and customizing existing technology [4].

There are several stages to the iterative research process: first 1) *defining* a problem, 2) *discovering and synthesizing* data, 3) *generating*, 4) *refining*, and finally 5) *reflecting* on solutions and evolving designs [4]. Through the course of this design process, RTD advances context, interpretation, and focus, and flexibly adapts applications to changing research goals and settings [4, 131].

The dual creation of an artifact and annotation are key to the method [4]; together, these outputs help users understand tacit knowledge embedded in design [41]. The final outcome of a design is a concrete problem framing, and a series of models, prototypes, and documentation of the design process [131].

3.2 Study Methods

The following section describes my study design and research methods. First, I describe the research methods I employed to understand expert and novice water monitoring practices (chapter 4). Then, I describe the research and design methods

I used to develop and test the StreamBED VR pilot training (chapter 5), and to test the role of multisensory cues on observation skills (chapter 6). Finally, I describe the methods I used to design the final StreamBED VR prototype (chapter 7), and to test its effectiveness against traditional PowerPoint (PPT) training.

3.2.1 Understanding Expert and Novice Training and Assessment

The goal of the initial study was to report on professional and volunteer stream monitor practices and to identify training opportunities and needs. This research was conducted using several composite metrics. I observed and shadowed expert monitors during assessment to understand their behavior, and conducted in-depth and group interviews with expert monitors. I also participated in citizen science training to understand group practices.

3.2.2 StreamBED VR Pilot Study

The StreamBED VR pilot study was an initial design and evaluation of stream training. The goal of this study was to consider the fit of immersive VR technology to qualitative assessment, and to identify user needs. The research was conducted as a usability study using mixed methods, pairing quantitative survey responses and scores with open-ended in-depth discussion. I triangulated quantitative analyses with qualitative thematic coding to understand how learners interacted with training, and to consider the challenges they faced.

3.2.3 Testing Effectiveness of Multisensory Cues on Observations

The multisensory study explored the role of training realism, a need identified by expert training and pilot feedback. The study employed several design and research methods. First, I used iterative quick-and-dirty prototyping [127] to create a multisensory system called the *Ambient Holodeck*. Using this system, I ran a between subjects study to compare how participants made observations and inferences of 360° stream habitat videos with and without sensory cues. I analyzed study data using content analysis to quantify participant observations, and paired it with grounded theory analysis to consider how participants made observations.

3.2.4 Testing StreamBED VR 2.0 against PowerPoint Baseline

The final study culminated my design research by testing two version of the StreamBED VR 2.0 prototype (Standard VR and Multisensory VR) against a PPT baseline. The final StreamBED VR training was iteratively designed based on findings from the previous three studies, using bodystorming and paper prototyping to develop the final prototype. Building off of the pilot and multisensory study, this research used mixed methods to understand how learners across three conditions interacted with and benefited from training. The study paired quantitative survey responses and participant scores with open-ended in-depth discussion, triangulating quantitative metrics with qualitative thematic analyses.

This chapter has described the research through design (RTD) method, and has outlined specific design and research methods I employed in my dissertation research. In the next chapter, I outline my first study, which considers how expert and citizen science water monitors learn to make qualitative stream assessments.

Chapter 4: Understanding Expert and Novice Training and Assessment

4.1 Overview

Engaging citizen scientists in substantive, high-level tasks has potential to improve data quality and participant retention [119, 124], however a traditional barrier is that developing the required skills is difficult and time consuming. Qualitative assessments are usually made by professionals who learn through iterative personal experiences. Learning to make qualitative assessments requires access to a variety of heterogeneous environments, which is often not practical for volunteers. In this chapter, I propose engaging potential citizen scientists motivated by learning goals with qualitative assessments of stream habitats.

This study considers the need for and feasibility of training citizen scientists to make qualitative assessments of streams and watersheds using a Rapid Bioassessment Protocol (RBP) [6]. To accomplish this objective, I observe and report on differences in background and training between professional and volunteer monitors, using these experiences to synthesize findings about volunteer training needs. My findings reveal that to successfully make qualitative stream assessments, volunteers need to: 1) experience a diverse range of streams, 2) discuss judgments with other monitors, and 3) construct internal narratives about water quality.

4.2 Method

The goal of this work is to assess the viability of training citizen science volunteers to make qualitative assessments of streams and watersheds using the EPA's Rapid Bioassessment Protocol (RBP) [6]. To compare current teaching methods employed by professionals and citizen science groups, I (1) observed and participated in RBP training and data collection with professional water monitors and with volunteers, (2) informally discussed water monitoring methods with ecologists, and (3) conducted semi-structured interviews with professional monitors.

4.2.1 Training and Data Collection

4.2.1.1 Professional RBP Training and Data Collection

I participated and observed water quality training with 4 professional water monitoring groups that included between 2 and 6 monitors. Each session lasted approximately 3 hours. Data collection took place at either 2 or 3 100-meter sites at different points of a stream.

In the larger teams, RBP assessment was paired with quantitative monitoring tasks. First, monitors measured the depth and velocity of the stream at different points, and measured quantitative measures such as temperature and PH. Then, they noted the presence or absence of stream characteristics like bedrock, and clay, and tallied the number of woody debris and roots in and around the stream. In addition, they collected samples of fish and "macroinvertebrates", water organisms like flatworms, crayfish and snails.

During each session, I observed how monitors made RBP qualitative assessments. Monitors first made personal evaluations, then compared their evaluations with a partner or group members, and settled on a numeric assessment for the metric. Monitors did this for all 13 RBP measures, then totaled them into an overall score. As well as observing professional monitors, I had a chance to experience their learning process firsthand; they made RBP assessments alongside other monitors, learned through group discussions, and performed assessments as part of a group (figure 4.1).

Figure 4.1: Left: Learning to make assessments in the field. Right: A group of water biologists discussing their observations and EPA protocol assessments.

4.2.1.2 Volunteer Training and Data Collection

I also observed and participated in volunteer RBP training.[1] Unlike professionals, who learn onsite, monitors were introduced to the RBP protocol through a 3 hour PowerPoint lecture [126] with images that exhibited a range of quality for the RBP stream characteristics (figure 4.2).

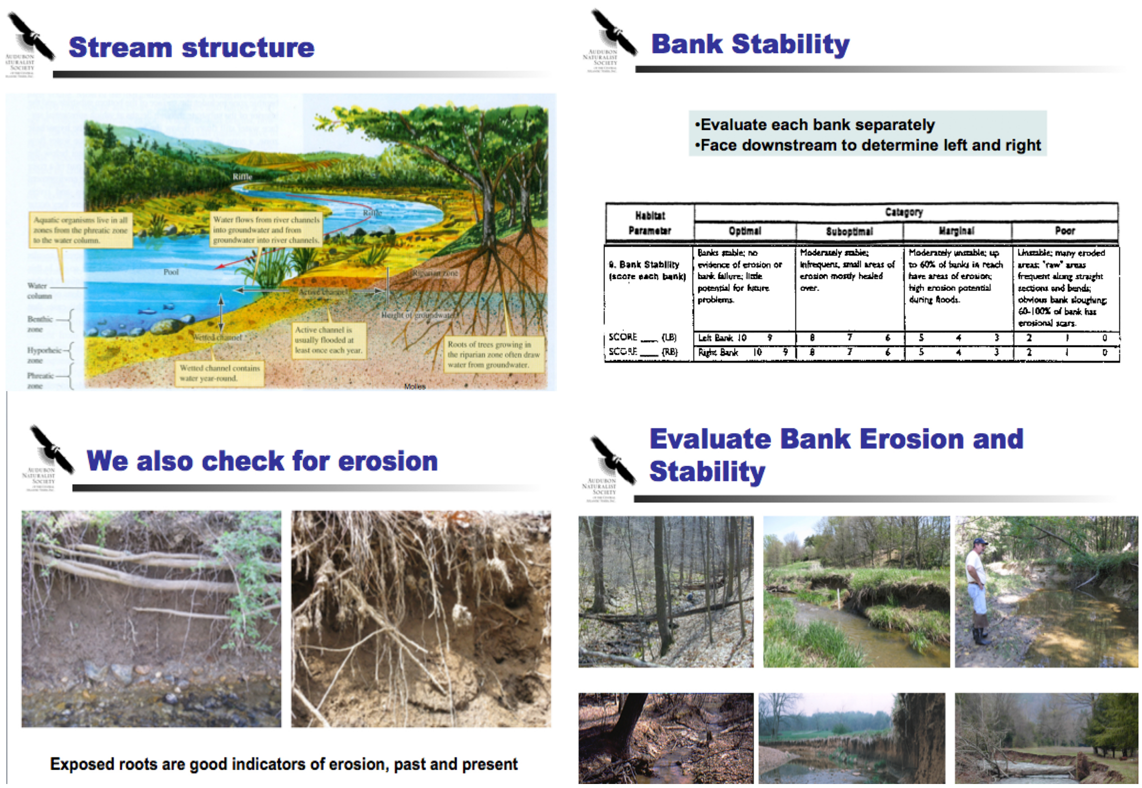

Figure 4.2: Examples of the RBP training slides. The top left slide overviews background information about the stream, the top right slide describes the "Bank Stability" protocol, and the bottom slides show images of streams with different bank stabilities.

As well as experiencing PowerPoint training, I participated in a 3 hour outdoor training and data collection experience (Figure 4.3) with 10 volunteer monitors.

[1] Although volunteers do not make qualitative assessments during data collection, they are introduced to the RBP scales as a background to water monitoring.

During data collection, monitors collected macroinvertebrates with hand-held fishing nets, separated them by species using a field guide, and counted them. After categorizing and tallying each species, volunteers practiced measuring the stream's PH and temperature.

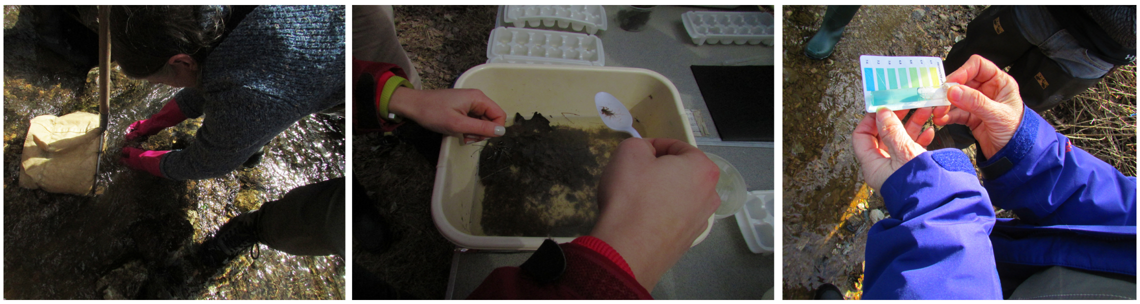

Figure 4.3: Images from volunteer outdoor data collection. During data collection, volunteers collected and counted macroinvertebrates, measured PH and temperature.

4.2.2 Informal Discussions with Ecologists

I received initial insights about the qualitative stream assessment process at the Association of Mid-Atlantic Aquatic Biologists (AMAAB), a regional conference for water biologists and ecologists. During the conference, I presented a poster describing early findings based on my in-person training experiences. During the 2 hour poster session, I discussed my findings with more than 15 ecologists, biologists, and water monitors. I also collected feedback on how the RBP protocol was employed by different monitoring groups, how my first-person training experiences compared to researchers' own training, and whether volunteers could be trained virtually to perform these tasks.

4.2.3 Semi-Structured Interviews with Professional RBP monitors

I conducted 5 in-depth interviews with professional monitors by phone or in person (after on-site training), each of which lasted approximately 1.5 hours. Participants included an aquatic biologist at the EPA, a program manager of the Virginia Department of Environmental Protection, and an ecologist at the Fairfax Department of Public Works and Environmental Services. Interviewees had between 2 and 23 years of experience making qualitative stream assessments, including one participant who helped develop and test the protocol in 1992.

During each interview, participants described (1) their personal process for making RBP assessments, (2) how they learned to make assessments, and (3) how they taught young professionals to make assessments (see Appendix A). In addition, I conducted an interview with the water quality monitoring program coordinator at at the Audubon Naturalist Society [83] to understand the nature of volunteer water monitoring tasks and training.

Several interview and feedback sessions were audio-recorded, however recording was not possible for all sessions, either due to a loud outdoor environment or a conference setting. When recording was not possible, I took detailed notes and followed up with participants for further discussion and clarification. I then listened to recordings, and transcribed important quotes. I chose not to transcribe recordings in their entirety because of budget and time constraints. However, I accounted for this by taking detailed notes during the interviews, intermittently stopping interviewees to repeat important words or phrases.

4.2.4 Data Analysis

I compiled notes, annotations, and quotes from the training and data collection experiences, informal discussions, and expert interviews into a single spreadsheet, and performed a thematic analysis of the interview data. The process took several stages of iteration. First a coder split the notes, annotations, and quotes into individual statements, and cleaned the data. Much of the data was collected in the field, so unrelated statements were discarded from the dataset.

I performed open-coding and thematic analysis [103] on the cleaned data, documenting patterns and ascribing color codes and labels to those patterns. During this time, I took notes on related quotes and reasoning behind open codes. Then, I created an affinity diagram of related open codes, identifying a set of larger themes. Using these larger themes, I refocused on the data, looking for quotes that supported my themes. I organized these new quotes under themes, and identified complementary sub-themes that emerged from the data.

4.3 Results

The following section compares current teaching methods employed by citizen science groups and professional water monitors, and describes challenges of making qualitative stream assessments. The section then describes salient themes from expert interviews and feedback sessions.

4.3.1 Differences in Background

I found that the professional monitors I met had a more extensive background in the natural sciences than the volunteers. All the professional monitors I interviewed either had a degree or had taken multiple courses in biology, ecology or conservation before focusing on water monitoring. In addition, professionals mentioned completing water-quality accreditations or certificates to learn to perform procedural tasks like fish and macroinvertebrate sampling. Although they had related training, the professionals I spoke to did not have a specific background in RBP assessment, they learned to do the qualitative assessment through first-person experiences.

In contrast to professionals, volunteers that participated in water monitoring training and data collection had a range of background experiences: some had degrees in biology or ecology, others had participated in other monitoring training, and a few had little to no experience in the domain, but were eager to learn. Notably, almost all of the volunteer monitors I spoke to during the citizen science outdoor training session had some sort of higher education background: several of the younger volunteers were recently out of college, and many of the older volunteers were retired educators. I do not have exact data on volunteer backgrounds because I learned about them informally, through conversations during outdoor data collection. However, finding that volunteers were highly educated is supported by previous work suggesting that citizen science attracts affluent volunteers motivated to improve society [105].

In addition, the Audubon society coordinator discussed several background courses that helped support volunteers with different backgrounds. Volunteers could learn about ecology through courses like *a natural history of aquatic ecology* and *healthy stream biology*. Likewise, volunteers could build procedural and identification skills through classes such as an *overview of invasive plants* and a series on *aquatic insect identification* [83]. Volunteers could even take certification exams to participate in certain projects or lead volunteer teams.

While professional monitors had more formal training than volunteers, a surprising finding is that there was less of a distinction between professional and volunteer monitoring backgrounds than I anticipated. Both groups had a higher education, had some experience in the natural sciences, and had opportunities to become better monitors through certifications. Notably, the biggest difference between professionals and volunteers was the number of streams each group had experience evaluating; professionals visited many more streams than volunteers, and were thus able to more easily compare them.

Since professional monitors had the opportunity to first-hand experience many more streams than volunteers, the way they learned about the qualitative RBP measures differed greatly from volunteers. Professionals learned to make assessments in an apprenticeship under more experienced monitors, whereas volunteers learned about the measures through a PowerPoint lecture that outlined the protocol measures, but did not transfer any nuanced or practical assessment skills.

4.3.2 Challenges Interpreting Qualitative Measures

As well as finding differences in background and learning between professional and volunteer monitors, my research uncovered multiple challenges in interpreting the protocol. The task of interpreting a protocol is quite similar in nature to interpreting a survey question; in order to make an informed response, a participant has to understand the meaning of the words in the question, understand the scale dimensions, and know how to map the question to the scale [40]. Similarly, in order to make an informed RBP assessment, an evaluator has to understand the contextual definitions of quality defined by each protocol metric, then consider what quality attributes correspond to scales values.

A primary challenge of the RBP is characterizing the state of an outdoor environment using a quasi-quantitative scale. To make an informed assessment, monitors must make subjective interpretations of multiple scale descriptions that are hard to quantify. Table 4.1 illustrates several interpretation issues that exist, including *interpreting scale measures*, *accounting for site variability, evaluating related measures*, and interpreting and *accounting for the passage of time*.

For instance, during a professional training session, monitors explained that a scale measuring embeddedness of particles in a stream bed (RBP protocol 2a [6]) should not be interpreted linearly, even though it was written to suggest linearity: the written scale suggests that 25% is suboptimal quality and 75% is poor quality. Professional monitors explained that realistically, more than 25% embeddedness should be characterized as poor quality because the environment becomes unsuitable for macroinvertebrates. Given this discrepancy, it is would not be clear to a volunteer monitor whether to evaluate 25% embeddedness as suboptimal or poor.

Similarly, some protocol measures describe quality using time measures that require heuristic interpretation. For example, in channel alteration (RBP protocol 6, shown in Figure 4.4), the suboptimal condition asks monitors to evaluate if there is evidence of channelization (stream straightening, widening, or deepening) "greater than past 20 years." This assumes that monitors can heuristically estimate the time frame of a disturbance. Further, this time measure is used asymmetrically, only to describe the suboptimal category. This increases the challenge of making evaluations because it is difficult to compare the suboptimal category to other categories.

From my experiences, I found that the subjectivity and imprecision of the RBP protocol compelled monitors to rely on background knowledge and personal experiences to make assessments. Notably, my informal conversations with ecologists revealed that the water monitoring community was aware of interpretation challenges that existed in the protocol, and made up for the protocol flaws with personal contextual knowledge.

Although they heavily relied on personal experiences, professionals suggested that their divergent monitoring experiences biased their assessment. For instance, an ecologist in Fairfax, Virginia worked with primarily urban disturbed streams, whereas a West Virginia monitor evaluated undisturbed rural streams. Due to their different backgrounds, the Fairfax ecologist was more likely to judge stream quality more leniently than the West Virginia monitor.

In line with these discrepancies, Roth [113] remarks that the monitoring com-

Table 4.1: Protocol Interpretation Questions and Challenges

How to interpret scales?	The RBP protocol suggests that measures should be assessed linearly, but professionals suggest linearity varies between measures. For example, professionals interpret a stream with 25% or more embeddedness as poor because the environment is unsuitable for macro-invertebrate organisms even though the protocol considers the habitat suboptimal.
How to Account for Site Variability?	Data collectors are asked to evaluate 100 meter stream cross sections, but how should users evaluate areas with significant variability? Professionals make holistic judgments of quality based on their experience, but new data collectors have no foundation with which to make such judgments.
How to evaluate related measures?	Several measures of stream quality directly affect one another (e.g. stream bank stability affects sediment deposits). How should data collectors account for this in their assessments?
How to interpret passage of time?	3 of 13 protocol measures ask users to evaluate transience of stream elements (e.g. logs and cobble) and recency of human activity (e.g. whether stream channel alteration occurred more or less than 20 years ago). How would users know how to judge the passage of time?

munity must resolve issues in field sampling protocols, differences in types of data collected, and discrepancies in condition ratings. Further, Roth appeals for increased accuracy of stream condition estimates.

4.3.3 Expert Interview Findings

As well as revealing the nature of professional monitoring experiences, the expert interviews uncovered multiple themes that illuminate professionals' qualitative assessment process. These themes illustrate how monitors make observations in outdoor stream environments, how they compare streams, identify patterns, and use these patterns to make RBP assessments.

Condition Category			
Optimal	**Suboptimal**	**Marginal**	**Poor**
Channelization or dredging absent orminimal; stream with normal pattern.	Some channelization present usually in areas of bridge abutments; evidence of past channelization, i.e., dredging, (greater than past 20 yr) may be present, but recent channelizationis not present	Channelization may be extensive; embankments or shoring structures present on both banks; and 40 to 80% of stream reach channelized and disrupted.	Banks shored with gabion or cement; over 80% of the stream reach channelized and disrupted. Instream habitat greatly altered or removed entirely.
20 19 18 17 16	15 14 13 12 11	10 9 8 7 6	5 4 3 2 1 0

Figure 4.4: *Channel Alteration*, a qualitative EPA metric that evaluates human impact on stream channels. This metric poses several challenges, including asking monitors to interpret the meaning of "*normal pattern*" streams and subjective time metrics (e.g."*greater than the past 20 years*"), and to judge percentages of channel disturbance.

4.3.3.1 Making Intuitive Judgments of Quality

The expert interviews revealed that professional monitors had a complex relationship with the RBP protocol. Professionals suggested that together, the 13 measures helped capture "a snapshot" of stream quality, but many agreed that the individual measures were either imprecise or challenging to interpret. This supports my experience of learning to make RBP assessments with professional monitors, and parallels my informal discussions with professionals.

I found that professionals with different amounts of expertise make RBP assessments very differently; less experienced monitors dutifully tried to interpret the protocol language, whereas more experienced professionals developed an intuition for assessment quality. Two interviewees, a senior biologist and ecologist, said they linked "mental images" from their experiences to the protocol scales. Likewise, the program manager who helped develop the protocol suggested that after making evaluations for 24 years, they could evaluate a stream at a glance, "without even scoring it." While the RBP scales are technically quantitative, expert interviews confirmed my theory that practically, the measures are qualitative.

4.3.3.2 Using Multisensory Information to Make Judgments

The water monitors and ecologists I interviewed described using multisensory environmental information to form opinions of a stream's quality. Several biologists mentioned supplementing the EPA protocol with additional measures for trash, presence of human activity, and invasive plant and animal species. For instance, an ecologist commented that hearing European Starlings (an invasive species) was indicative of poor stream quality. Likewise, another noted that invasive plant species often indicated that a stream habitat has been disturbed. Although these heuristics were not formally part of the qualitative assessment protocol, professionals suggested that paying attention to additional habitat characteristics helped them identify environment stressors that affected the RBP measures.

Professional monitors also used ambient sensory information to discern habitat stressors. For example, one ecologist approximated the strength of stream riffles by the sound of rushing water, whereas another used sun warmth and wind strength to judge the density of stream bank vegetation. Since formal assessment was limited to

100 meter cross-sections, monitors conferred that additional sensory information allowed them to make more comprehensive observations about the state of the stream that they may not have readily been able to see. This is in-line with Dede's 1999 findings [27], that multisensory information helps students understand complex scientific models through experiential metaphors and analogies. This is likewise in line with Dede's (2017) [29] work on stream identification tasks, finding that sensory information (*e.g.* sound, color and turbidity of the water, weather variables, shifts in grass color) helped learners sense pattern changes.

4.3.3.3 Describing Stream Quality using Past, Present, and Future Narratives

As well as observing sensory information and making intuitive judgments, several monitors I spoke to actively interpreted "narratives" of the stream spaces: how the landscape had transformed from the past, and how the stream in its present state would shape its future. For instance, during an on-site learning experience at a stream in northern Virginia, a professional monitor pointed to markers indicating that a stream was part of a historical agriculture site. They explained that the former use of the land had caused extensive erosion at the monitoring site by irrigating water through the stream, causing faster moving water that wore away at the stream's banks. Using the stream's history and current state, the monitor predicted that the stream would become wider and more eroded in 5-10 years.

Likewise, during an expert interview, a biologist emphasized how connected ecosystems were, and characterized stream landscapes as a consequence of events; for instance, they recounted that sedimentation was often caused by homeowners mowing a stream's vegetative zone to keep snakes off of their property. However, they explained that removing bank vegetation removed roots that held soil in place, and that the whole bank would erode the next time the stream flooded. Rather than merely evaluating streams in their current form, professional monitors predicted what caused stream characteristics, and how those characteristics changed over time.

4.4 Summary

The overarching goal of my work was to assess the viability of training volunteers to make different types of inductive qualitative assessments. To undertake this challenge, I used the domain of water monitoring to understand how professionals train and make qualitative assessments, and how volunteers learn about qualitative assessments. I observed and participated in RBP training and data collection with professional water monitors and citizen scientists, informally discussed water monitoring methods with ecologists, and conducted semi-structured interviews with professional monitors.

I found that professionals learned to make assessments on-site, through iterative assessment and discussion with peers and instructors. Through onsite experiences, I found that professionals develop intuitive judgments of quality, use multisensory environmental information to make judgments, and construct past and future narratives of streams using environmental characteristics. I also found that the qualitative RBP protocol is subjective and misleading, perhaps because it tries to quantify intrinsically qualitative measures.

Contrary to my expectations, I found that volunteers primarily differed from professionals in the number of streams they had visited and assessed. To match professional training experiences, I identified 3 training needs; to first-hand experience environments in order to develop intuitive judgments, to discuss judgments with other monitors, and to form narratives of stream quality from assessments.

Chapter 5: StreamBED VR Pilot Study

5.1 Overview

The goal of my research is to give citizen scientists the same interpretive skills as water quality biologists, so that they they can make appropriate qualitative judgments. In this chapter, I present a new training system, called StreamBED VR, that teaches volunteers to qualitatively assess virtual streams by physically exploring and inspecting them from multiple angles and perspectives. I also describe the StreamBED VR pilot study, an initial design and evaluation of the system.

The goal of this study was to consider the viability of the StreamBED VR initial training design, and to understand novice learner needs. To accomplish this goal, this study taught participants who are not expert monitors to make and update qualitative assessments of 4 EPA protocol metrics [6]. During training, participants first saw differences in habitat quality by experiencing an "optimal" and a "poor" habitat stream, then learned to calibrate assessments by evaluating and getting feedback on virtual streams exhibiting diverse quality characteristics.

My findings reveal that 1) the StreamBED VR training immersed and motivated study participants, inspiring them to 2) interact with the environment through storytelling and virtual surveying. However, pilot findings revealed that 3) participants had trouble with several facets of the training; they were distracted by the rendered training environment, had trouble interpreting protocol subjectivity, and anchored to extreme protocol judgments.

In the following sections, I introduce my StreamBED VR training system, then present my study method. Next, I describe my quantitative and qualitative thematic findings. Finally, I summarize my findings about the viability of VR training and participant challenges and needs.

5.2 Training System Design

The following section describes the StreamBED system design. First, I describe the Oculus Rift and Unity platforms, and describe the protocol measures used in the training. Then, I describe the "optimal" and "poor" tutorial environments, and outline tutorial interaction and assessment tasks.

Figure 5.1: Pilot study participant exploring the virtual environment with an Oculus Rift and Xbox 360 controller.

5.2.1 Oculus Rift and Unity Game Engine

Unity is a game development engine that supports virtual reality integration. The training environment was developed in Unity 5 and integrated with the Oculus Rift HMD. The virtual training streams were constructed to reflect the EPA's Bioassessment Protocol Guidelines [6], using brushes, textures, assets and prefabs found in the Unity Asset store and online. In order to simulate realistic interactions, participants interacted with the training environment using an Oculus Rift SDK2 and Xbox 360 Game Controller, as shown in Figure 5.1.

5.2.2 Protocol

Four measures from the Bioassessment Protocol [5] were simulated in the Unity virtual environment: 1) Epifaunal Substrate 2) Bank Stability, 3) Riparian Vegetation Zone Width, and 4) Channel Alteration. These measures were chosen for their considerable impact on habitat (based on correlations to biological index scores) [6] and relative difficulty to understand and interpret based on questions described in Table 4.1. The measures were modeled on RBP protocol guidelines [6].

5.2.3 Optimal and Poor Tutorial Environments

Training consisted of participants navigating two tutorial environments: an *optimal quality* stream featuring an abundance of epifaunal substrate, a large riparian zone, stable banks without channel alteration, and a *poor quality* stream with no epifaunal substrate, no riparian zone, high channel alteration, and highly eroded unstable banks. In addition to modeling these measures in the training environment, additional measures of stream quality were built in, including variability in pool depth, water clarity, and diversity of vegetation; this was designed to make the experience realistic, and to challenge participants to make assessments in context of other factors.

5.2.3.1 Tutorial Interaction

Participants first made topical observations of quality as they walked through the tutorials, then picked up definition and protocol cards around the training sites

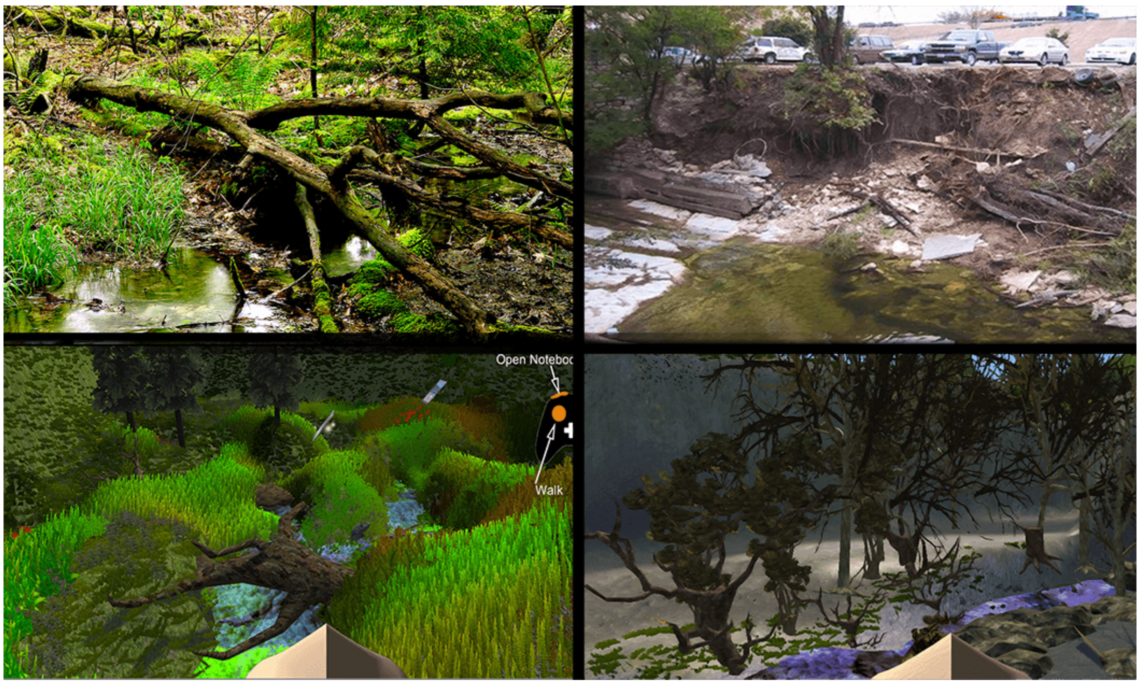

Figure 5.2: Left: Optimal stream environment, as shown in a photograph (above) and simulated in the StreamBED training (below). The optimal environment included stable banks covered with vegetation, a large riparian zone, no channel alteration and a lot of epifaunal substrate (made from fallen logs, cobble, plant vegetation, and undercut banks). Right: Poor stream environment, shown in a photograph (above) and simulated in the StreamBED training (below). The poor environment showed unstable, "raw" soil with high evidence of erosion and bank failure, no riparian zone, and no epifaunal substrate.

(Figure 5.3), and added them to a virtual field notebook (Figure 5.4). The definition cards contained definitions for technical or unfamiliar words used in the protocol. In the optimal environment, the protocol cards displayed optimal images and descriptions of the 4 bioassessment measures, and in the poor environment, the cards displayed poor images and descriptions of the same measures.

The goal of adding protocol cards and definitions to a notebook was to simulate the experience of collecting observations into a field notebook. After participants picked up the protocol and definition cards, the text and images on the cards appeared in the notebook, building a reference for participants to consult during assessment tasks. Participants later used to this virtual reference notebook to make virtual stream assessments.

5.2.4 Assessment Task

After completing the optimal and poor tutorials, participants explored and evaluated 3 virtual streams with a diverse range of quality characteristics (Figure 5.5). Participants evaluated all 4 protocols in each of the stream environments, making a total of 12 assessments.

The goal of virtual assessment was to give learners the opportunity to practice stream assessment by surveying the stream landscape and referencing the virtual notebook. Participants made assessments by picking up "assessment cards," that had the name of a protocol and a scale on it (Figure 5.4). Participants made assessments by moving the slider on the card to a number corresponding to the

Figure 5.3: Two protocol information cards. Left: A definition card for the channel alteration protocol. The card defines the term "shoring structure." Right: An "optimal" protocol card for the channel alteration protocol. The card shows an example of an optimal channel, and defines the optimal condition below the image.

Figure 5.4: Left: A virtual field notebook. Participants "collected" definition and assessment cards into this notebook, then used it as reference during the assessment tasks. Right: an assessment task card for channel alteration. Participants first took a photo of the stream, and then rated it on the Channel Alteration 20-point scale.

protocol scale.

It was important to learn from mistakes. For this reason, participants that make incorrect assessments were prompted to reevaluate their assessment of each protocol until they scored within 2 points of a predetermined "correct" response.

5.3 Method

Ten (10) participants (6 men and 4 women) were recruited from a pool of graduate students taking classes at the University of Maryland who successfully passed a simulator sickness prescreening, and who consented to participate in the study. All 10 study sessions were audio-recorded.

The study consisted of a VR training session and an assessment of a stream on the University of Maryland campus. During the study, participants interacted with the training using an Oculus Rift SDK2 and Xbox 360 Game Controller with

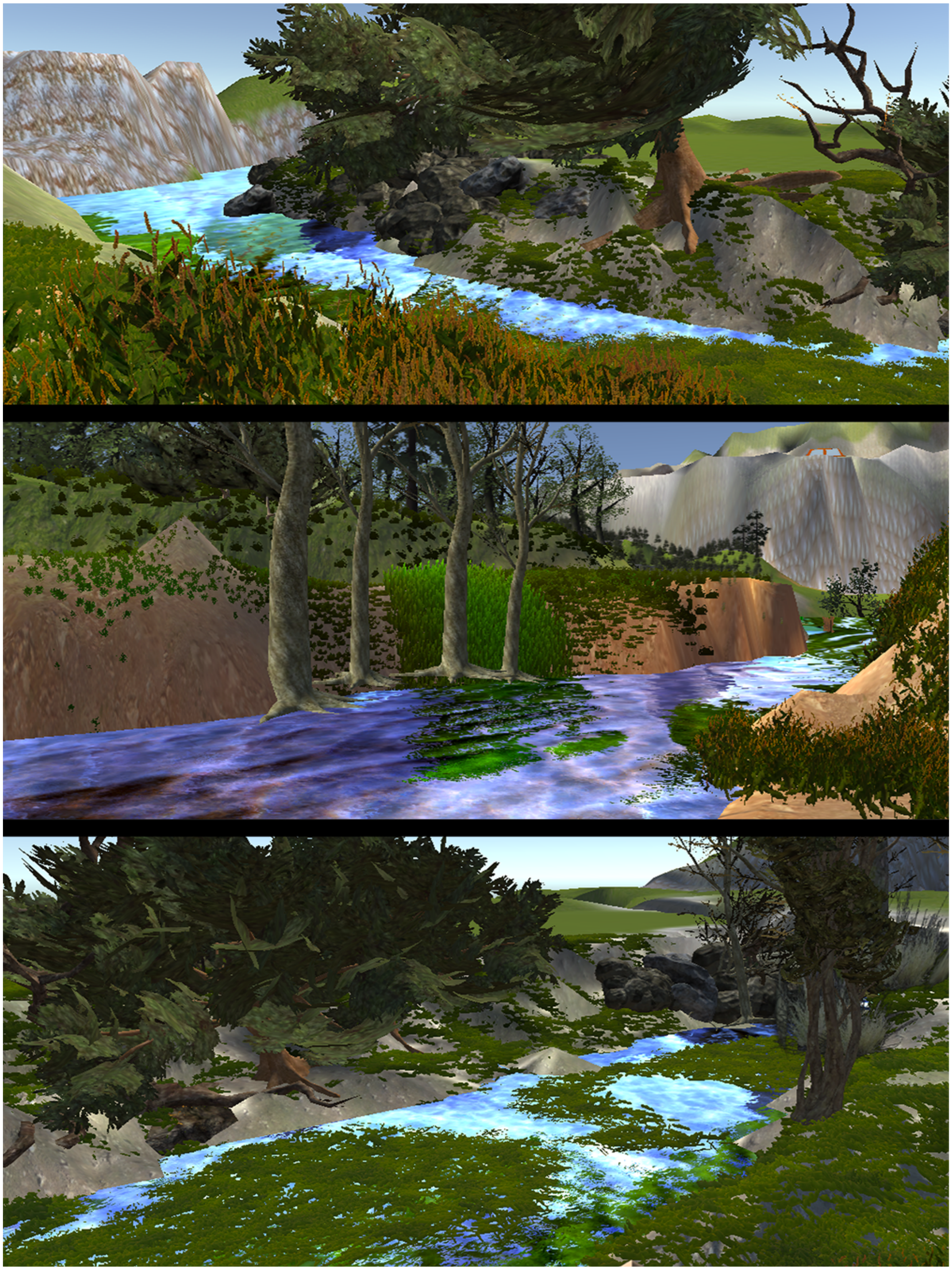

Figure 5.5: Screenshots of the three assessment environments showing a range of epifaunal substrate, bank stability, channel alteration, and riparian zone characteristics.

standard key mappings.

5.3.1 Training

After signing consent forms, participants filled out a background questionnaire assessing their demographics and background variables, their experience with citizen science, and with water monitoring. Before the study, participants were also asked to rate their excitement for the study, and to predict their engagement during training.

After filling out the questionnaire, participants were introduced to the topics of citizen science and water quality monitoring, and were guided through a short Xbox controller training that allowed them to practice navigation while seeing the controller. After controller training, participants put on the Oculus Rift HMD and completed the training tasks described above. During the VR tasks, participants were provided with water and were encouraged to take breaks when they felt uncomfortable. Those who had trouble wearing the Oculus Rift (primarily due to glasses) visually explored the tutorials using the Oculus Rift, but completed card collection and assessment tasks using a high definition monitor. After training, participants completed a post-training survey assessing the usability of the VR environment, and responded to questions about training engagement and environment presence.

5.3.2 Outdoor Assessment

After training, participants assessed a stream on the University of Maryland campus. During the walk to the stream, participants evaluated their experience

Figure 5.6: Pilot participant making an assessment at a stream at the University of Maryland. Positive features included a mix of snags, logs and cobble (epifaunal substrate), soft muddy eroded banks (bank stability), a straightened stream shored with artificial gabion rocks (channel alteration), and a small riparian zone (riparian zone).

with the virtual training, and predicted their ability to accurately assess the measures they had learned. At the stream, participants received a physical copy of the virtual reference notebook, and were asked to rate the stream's features based on the measures they learned, orally explaining their reasoning and decisions (outdoor data collection shown in Figure 5.6). After completing the outdoor assessment, participants filled out a final questionnaire about the relevance of the virtual training to the physical task, and answered questions about their confidence and motivation to participate in future water monitoring projects.

5.3.3 Analysis

Study analyses triangulated quantitative survey responses with qualitative thematic coding of audio-recorded discussion. First, I describe the performed quantitative analyses, then describe the qualitative thematic coding process.

5.3.3.1 Quantitative Metrics

The small study population (n=10) limited the quantitative analysis to descriptive statistics, correlations, and scatterplot trends. Through these analyses, I show shifts in participant immersion, motivation, and self confidence of the course of the study. All survey data was measured on a 7-point Likert scale.

5.3.3.2 Thematic Analysis

Qualitative audio-recording data was analyzed using thematic analysis. First, salient quotes from participant audio-recordings were transcribed and open-coded to identify codes about user experience during training and data collection. Then, through several stages of grouping and regrouping codes, I created an affinity diagram of related open codes, and identified a set of larger themes. These identified themes were used to assess the effect of the embodied training design on participant enjoyment, immersion, and motivation throughout the study.

5.4 Results

5.4.1 Background Variables

The study collected background data on participant demographics and background in education. Seven of 10 participants were familiar with at least one citizen science project and had participated in data collection. Participants were on average not very familiar with water monitoring ($\mu = 1.88$ of 7), and none had participated in water monitoring training or data collection.

Of collected background variables, self-efficacy, participants' assessment of their ability and confidence to complete tasks and reach goals, was highly correlated with self-prediction variables, including the expectation to enjoy training (ρ=.894), and data collection (ρ=.904). Self-efficacy was likewise correlated with outcome variables, including training engagement (ρ=.773), presence, immersion (ρ=.584), assessment task confidence (ρ=.767), and likelihood to participate in future data collection (ρ=.458). The effect of self-efficacy on motivation has been well documented [68, 5] as a driving force of engagement and immersion. This data is consistent with previous research; participants who were confident in their ability to perform assessments may have enjoyed training and testing tasks more than participants who were not confident.

5.4.2 Study Outcomes

Study participants completed virtual training and assessment activities in approximately 2 hours and 10 minutes: training took an average of 2 hours (although there was a large spread in the data due to individual differences and technical errors), and outdoor data collection took approximately 10 minutes.

Participants spent an average of 56 minutes picking up cards and exploring the optimal and poor environments, and approximately 62 minutes making assessments of the virtual test stream. During the virtual assessment task, participants took 14.47 tries to answer the 4 assessment questions, and gave an average of 2.64 incorrect responses for every assessment question, with a large spread between 1 and 6.5 incorrect tries per assessment task. Although participants had a large error rate, the assessment task was as much a learning tool as the optimal and poor tutorials; during the assessments, it took time for participants to associate protocol scales with virtual environment features. It is thus not surprising that participants had high error rates given the learning curve.

After training, participants made protocol assessments at an outdoor stream on the campus. The outdoor assessments were on average 2.37 points away from the correct response, a "gold standard" assessment made by the researcher and vetted by a water quality biologist [101]; differences between participant and gold standard scores were calculated by taking the absolute value of the participants' score subtracted from the gold standard score. Participants were an average of 3.25 points away from the correct response on protocol scales ranging from 0 to 20, and were on average 1.93 points away from the gold standard on scales ranging from 0 to 10.

Notably, there was a relatively strong positive correlation between total virtual training time, and participant's outdoor data collection scores; participants who spent more time on training also made assessments that were closer to the outdoor gold standard. Strikingly, there was a negative correlation between the amount of simulator sickness participants experienced, and the amount of time they spent on training (ρ=-.338). Participants who felt greater simulator sickness did not spend as much time training, and also did not collect as accurate data as participants who felt mild or no sickness.

5.4.3 Immersion and Motivation

After VR training, participants answered questions about system immersion and usability. Participants also rated their immersion and motivation to perform training and data collection tasks throughout the study (Figure 5.7). They assessed their immersion before and after training, and rated their motivation to collect data that day and in the future. Average usability for participants ranged from 3 to 6.5, and average immersion ranged from 2.5 to 6. There was a large positive correlation (ρ=.706) between system usability and participant immersion.

Figure 5.7 demonstrates the significant shift in immersion from the beginning to the end of the study: at the beginning, average participant expectation to enjoy training (a proxy for immersion) centered at $\mu = 4$, after training shifted to 5.5,

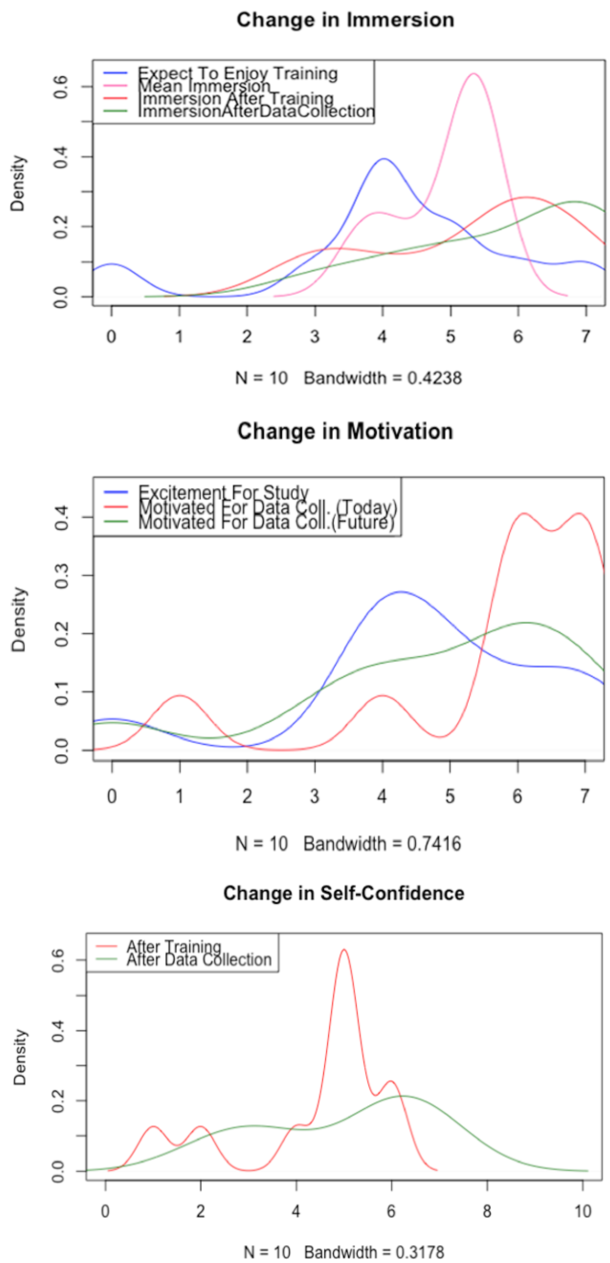

Figure 5.7: Three scatterplots showing how participant immersion, motivation, and self confidence changed over the course of the study. Top: Scatterplot showing participant expectation to enjoy training (blue), mean training immersion (pink), immersion after training (red), and immersion after data collection (green). Middle: Scatterplot showing change in participant motivation over time; excitement for study (blue), motivation for outdoor data collection (red), motivation for future data collection (green). Bottom: Scatterplot showing change in self-confidence over time; after training (red) and after data collection (green).

and after data collection shifted all the way to 7. Figure 5.7 also demonstrates participants' shift in motivation. At the beginning of the study, average excitement (a proxy for motivation) centered at $\mu = 4$. After training, average motivation to collect data shifted to approximately 6.5, and after data collection shifted to 6. These two trends are consistent with a positive effect of training and assessment tasks on immersion and motivation. The shift in immersion and motivation is paralleled in participants' assessment of their confidence to collect meaningful stream data: before to after data collection, confidence shifted from $\mu = 4.5$ to $\mu = 6.5$.

5.4.4 Qualitative Themes

The qualitative evaluation of StreamBED strove to understand whether the training could meaningful teach citizen scientists to make holistic stream assessments. Several themes emerged from the open coding data: Participants 1) *told stories* to make sense of environments, and 2) used *virtual training affordances to survey* the environment. Trying to make sense of the stream habitats, participants also 3) *considered extraneous information* during assessments, 4) *judged moderate streams using extreme standards* and 5) *struggled with protocol subjectivity.*

5.4.4.1 Stories as Information

Participants created stories that helped them explain virtual phenomena. One person joked that "*some crazed lumberjack came through here...and was...so buff that he picked up the logs with his hands and walked off with them.*" Likewise, and another participant noted that "*human activity screwed the stream up so bad that it sort of grew back.*" Others saw the story of the changing landscape; "*There's not enough water here...some of the land that used to be the waterbed now [has] no water.*" Another participant used the landscape to evaluate monitoring features; "*the rock around the channel tells me that this may be unnatural...everywhere else there is no rock.*"

5.4.4.2 Virtual Surveying

The virtual environment afforded participants with versatility that allowed them to meaningfully understand the environment. During an assessment task, one participant said that she was "*going to the higher ground to...have a better look,*" and several commented on their ability to be underwater, something hard to do in the real world. One participant noted that they could "*explore in a really intuitive way...I can walk real fast and look around in a more efficient way without getting muddy.*" After training, a participant even commented that the VR training was more engaging than actual data collection; "*its similar,*" they said, "*but I cannot do as much as I could in the VR.*"

5.4.4.3 Using Extraneous Information

Participants were directed to make evaluations using only protocol features, however they often took superfluous features under evaluation. Several participants commented that the water in the poor environment "*seems a little artificial*" and the grass in optimal environment ‘*‘was a little too green.*” One participant asked about the variety of plants that were growing along the stream bank; “*[are these] supposed to be rice? Aquatic corn?*” Likewise, a participant commented than an environment looked poor because there weren't “*many leaves on [the trees].*”

5.4.4.4 Extreme Standards

After seeing the optimal and poor tutorials, participants frequently judged moderate streams using extreme language. Rather than describing habitats as "sub-optimal" or "marginal," one participant described a test environment as “*closer to poor than optimal,*” while another said one area was “*optimal...the left bank...but in the poor environment its something like...this side without grass.*” Although the tutorials biased participants toward extremes, their answers suggested that they were using their holistic experiences to guide the evaluations. For instance, a participant remarked, “*I won't [say] poor because there's not too much human being activities...it may be suboptimal...because it just don't look like...optimal.*”

5.4.4.5 Protocol Subjectivity

Participants also had trouble with the subjectivity of protocol descriptions and numeric mappings. For instance, one participant did not understand the differences between protocol subheadings. “*Why can't you choose 6 instead of 8 when they're both suboptimal?*” they asked. Similarly, another participant assumed that they were supposed to first choose a category (e.g. marginal) before choosing a numeric answer. Yet another participant wondered if the scores were percentages. "*6 means 60% and 7 means 70%?*" they asked, incorrectly assuming that measure scores mapped directly onto percentages. Even when participants understood the protocol, they commented on its subjectivity. A participant trying to interpret the protocol commented that it "*had weird mappings...[that] felt unnatural,*" and wondered "*why can't you do 4 levels, like optimal, suboptimal, marginal, poor?*"

5.5 Summary

The goal of this study was to consider the viability of VR to recreate outdoor stream habitats, and to consider participant training needs. Quantitative study findings revealed a positive shift in immersion and motivation, and qualitative themes suggested that participants went out of their way to interact with the training by developing stories while surveying the habitats. Together, the quantitative and qualitative findings suggest that the VR training created a dynamic learning environment that engaged learners and motivated them to understand the virtual habitats through physical exploration.

Table 5.1: Pilot Study Challenges and Design Questions

Challenges	Design Questions
Improving Habitat Realism	*How can simulated textures be made to look realistic? Is there a way to create photorealism in VR? Should habitat simulations include indirectly relevant ambient features?*
Interpreting Protocol Metrics	*How to help participants interpret subjective protocols? How to help participants focus only on relevant features?*
Scaffolding Learning Process	*What actionable goals and tasks can guide participant interaction? How to account for differences in background and experience?*
Improving Interaction Experience	*How to streamline and chunk training? How to give meaning to learning and assessment tasks? How to reward participants for learning?*
Giving Appropriate Feedback	*What kind of feedback best supports learning? Should scaffolding and feedback be provided by the system, an expert, or a peer learner?*

Study findings revealed several interpretation challenges, described in table 5.1; many participants did not fully understand the protocol scales, and required clear guidelines to interpret and compare measures to assessment habitats. Further, participants had trouble interpreting simulated VR habitats, and focused on irrelevant environment details, such as the color of trees. This suggests that participants required habitat realism during training, and needed concrete guidance to focus their attention only on relevant habitat features.

The pilot likewise revealed interaction challenges, also described in table 5.1. One challenge was that participant training requires clear training goals and tasks even though exploring and evaluating a habitat is a passive task. Training thus needs to consider what active tasks could guide participants through the training process. A second challenge was scaffolding training for individual participant needs, which is difficult because citizen scientists have a wide range of backgrounds and experience. Finally, a third challenge was to give participants meaningful feedback and rewards that support their learning process.

Chapter 6: The Effect of Multisensory Cues on Habitat Observation

6.1 Overview

This chapter considers the role of multisensory realism on participant observation skills. The two previous studies found that water quality experts employed multisensory cues in their assessments, and that pilot study participants required realism to make habitat assessments. In the context of citizen science water quality assessment training, this study asks whether giving potential citizen scientists additional sensory information in VR helps them make more detailed observations and inferences, two key components of making qualitative judgments [119].

This study first build a new multisensory environment interface for VR, called the *Ambient Holodeck*, that allowed users can feel and smell landscape environmental conditions (wind, temperature, humidity). I used this system to compare how participants made observations and inferred patterns between 2 stream habitats in VR, in either Standard VR with audio-visual cues, or in Multisensory VR with ambi-

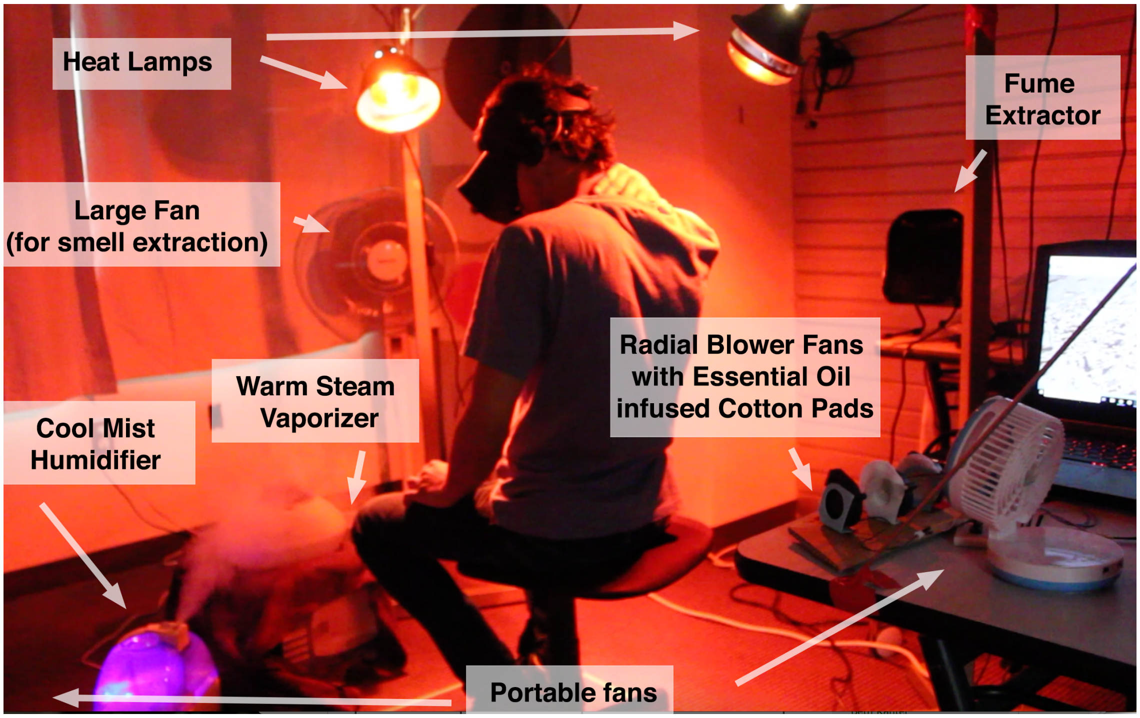

Figure 6.1: The multisensory environment setup. To create a flexible ambient environment, I used 2 heat lamps, a warm steam vaporizer, a cool mist humidifier, and 2 portable mini fans. I also developed a smell apparatus, described in Figure 6.2. After each study, a large fan and fume extractor removed scents from the room.

ent sensory cues. My findings reveal that multisensory information 1) improved the number of observations participants made, and 2) positively impacted engagement and immersion.

In the following sections, I introduce my multisensory system, then describe my experimental setup and analyses. Next, I describe my results about pattern recognition between Standard and Multisensory VR groups. Finally, I summarize the implications of these findings.

6.2 Designing the Ambient Holodeck

Below, I overview the design of the Ambient Holodeck Multisensory system by modality. In each section I first describe my design goals, overview challenges I faced, and describe my design process. At the end of each section, I present the final design for each sensory component.

6.2.1 Olfaction

I wanted users to experience environmental "smellscapes." It was important for olfactory elements to be distinguishable from one another so that they genuinely represented the kind of conditions that participants would experience at the different habitats. It was also important to distinguish them from the general study space.

Challenges: I identified 2 design challenges. One challenge was to dissipate environmental smells between studies quickly without bulky equipment. Another challenge was to make the olfactory experience ambient without seeming artificial or directional.

Design Process: I explored multiple designs during my process. First, I designed a rotating wheel mounted onto an Oculus head-mounted display (HMD) with cotton pads next to the participants' nose. This system was passive and expected users to take deep breaths. I abandoned this system because the powerful scents made pilot participants uncomfortable, and weighed down the headset. My second prototype used a desktop fan to blow air towards the participant through cotton pads charged with essential oils. This design was ambient and more comfortable,

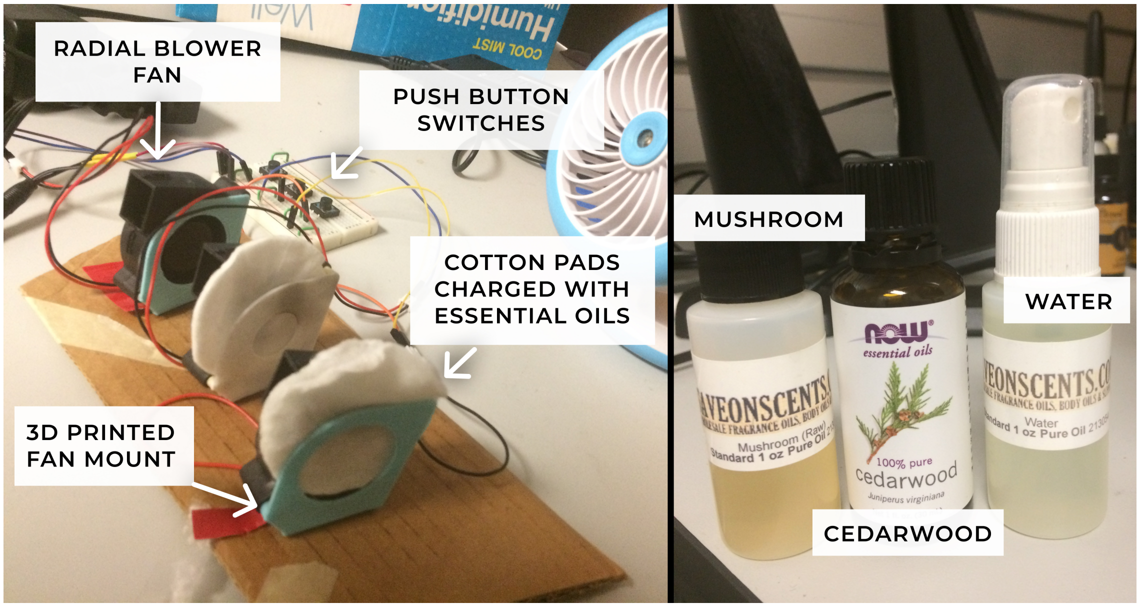

Figure 6.2: Left: Overview of the olfactory design. Radial blower fans were mounted with essential oil charged cotton pads, and activated using push buttons. Right: The scents I used for the study. For the hot spring, I used a muggy humid oil called "water," and for the redwood forest, I used a combination of cedar and mushroom oils.

but created appreciable wind that overpowered the fans I was using to simulate wind.

Final Design: The final design used three 5v 5015 cooling blower fans mounted with their exhausts facing the participant. Two cotton pads charged with scents were placed on the intake sides of two fans using a 3d printed mount. A third fan blew fresh air towards the participant to remove lingering scents. The fans were powered by a 5v power supply, and were connected to 3 push button switches. During the study, the researcher switched the fans on for a few seconds at a time. Between studies, lingering scents were cleared using a floor fan and fume extractor. The fan pushed the scented air across the study space into the fume extractor.

6.2.2 Thermal

The goal of the thermal system was to realistically replicate directional and ambient heat from the sun.

Challenges: I encountered two challenges in recreating ambient heat. The first challenge was to find a powerful and safe directional heat source. A second challenge was to find a powerful heat source that could quickly heat the study space.

Design Process: I considered several heat sources during my design process. First, I tested a personal space heater (Lasko #100 MyHeat Ceramic Heater) and large space heater (Mainstays Oil Filled Electric Radiant Space Heater). I found the personal heater provided reasonable heat up to two feet from the participant, but it became ineffective when it was placed any further. Likewise, the large space heater took upwards of 20 minutes to heat the room appreciably, rendering it ineffective for my purpose.

My second design mounted hair dryers on poles pointed towards the center of the study space to provide directional and ambient heat. The hair dryers were effective, but they generated considerable wind and noise that could be heard through the VR headset.

Final Design: My final design used two Philips 250W incandescent heat lamps mounted in reptile lamp stands. I configured these lamps to the direction of the sun in my 360 ° videos; one lamp simulated the sun, while the other provided ambient heat. The first video was of a sandy open environment (Figure 6.4). To simulate this habitat, both lamps were switched on, providing the participant with directional heat from the sun and ambient environmental heat. The second video was of a shady forested habitat. For this video, only the overhead heat lamp was used to provide ambient heat.

6.2.3 Humidity

In addition to olfaction and heat, it was important to realistically simulate the evaporated moisture created by diverse stream habitats. In the first video, I needed to replicate the humidity of a hot spring, and in the second video, I needed to replicate the cooler ambient humidity of a redwood forest stream.

Challenges: I encountered three challenges in generating ambient humidity. One challenge was to generate enough water vapor to simulate humidity. A second was to distinguish warm humidity from a cooler mist. A third challenge was to make the humidity feel directional, as though it was coming up from a stream.

Design Process: I considered different humidifiers and placements during design. The first design using two portable Cingk personal cooling misting fans, placed by participants' hands and feet. I found that by themselves, the misting fans were too small to generate ambient humidity. Further, fans had to be placed next to the participants, and were prone to being knocked over.

In my second design, I placed a Vicks Pediatric Steam Vaporizer on the floor. The vaporizer generated steam, simulating the hot spring in the first habitat. However, the vaporizer was uncomfortably hot when placed close to participants' feet, and could not provide the cool mist I required for the redwood habitat.

Final Design: My final design combined the vaporizer with a Walgreens Cool Mist Room Humidifier. The hot vaporizer and cool humidifier were angled toward each other, approximately two feet from participants. Combined, the humidifier provided participants with a strong directional mist, while the vaporizer generated ambient steam around the study space. In the first video, both the vaporizer and humidifier were switched on, mixing hot and cold mist. In second video, only the cool humidifier was switched on, simulating directional mist rising up from the forest stream.

6.2.4 Wind

My goal was to generate wind that felt realistic to participants. In the redwood habitat, I needed to simulate a gentle cool breeze.

Challenges: My biggest challenge was finding a source of wind that did not feel artificial and did not drown out the VR audio.

Design Process: I tested different size fans to simulate the breezy forest in the second video. First, I tested an oscillating floor fan, but found that it was too strong and loud to simulate a gentle breeze, and was immediately recognizable as a fan. Then, I tested two battery-powered Cingk personal fans at different speed settings in different configurations around the study space.

Final Design: The final design used the personal fans on their lowest speed, facing the participant from opposing directions. From a few feet away, the fans provided unobtrusive ambient wind that was not immediately recognizable as a fan.

Further, the fans' portability allowed them to be easily hidden between studies.

6.2.5 Environment Space

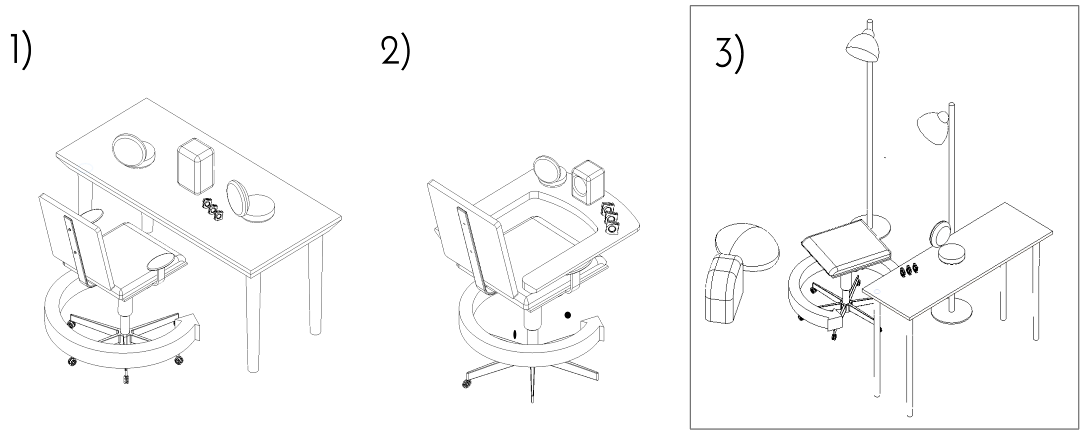

Figure 6.3: An overview my 360° environment design process. The final setup is shown on the right.

My environment space goal was to allow participants to rotate freely in a 360° spherical video. It was important for participants to experience the directional Multisensory components from different angles, and to extend their arms in order to point and gesture.

Challenges: My primary challenge was to allow participants to experience ambient sensory information from all directions.

Design Process: As I explored ambient sensory designs, I tested complementary ergonomic designs (see Figure 6.3). My first design placed sensory components on a desk in front of participants, however I found that they were too small and far away to create realism. My second design attached a removable tray to a full back swivel office chair, and placed sensory components on this tray. The tray was constructed from foam-core and was held onto the chair with Velcro. This light and sturdy design allowed users to freely turn the chair around to explore the space in VR. The chair was easy to assemble, but required participants to have their hands confined to the armrests in order to not tip over the equipment. Further, equipment was mounted to the front of the chair, so participants could not feel the ambient sensory information behind them.

Final design: My final design placed sensory components around the study space, and replaced the office chairs with a backless swivel chair. To prevent headset tangling, I ran the HMD wires through a hollow aluminum pole above the participant's head. This allowed participants to to spin freely.

Shown in Figure 6.1, the final design integrated the olfactory, humidity, thermal, wind, and ergonomic designs. Study participants were seated in the middle of the multisensory space on a backless swiveling office chair. Participants felt environmental warmth from the two mounted heat lamps, and felt hot and cool humidity from the cool mist humidifier and the steam vaporizer. In complement to this, participants felt breeze through two portable fans. Finally, participants smelled the environment through an olfactory mechanism I developed using radial blower fans (Figure 6.2).

6.3 Method

The goal of this work is to consider how multisensory realism affects how amateur water monitors make observations and recognize watershed habitat patterns.

To consider the effect of mulsemedia on pattern recognition, I: (1) asked participants to make observations and inferences of two watershed habitats, either in 1) Standard VR with audio-visual VR experience, or 2) in Multisensory VR with additional sensory landscape and environmental conditions. I performed a content analysis of these observations and inferences to discern the levels of detail with which participants described the environment, the sensory information they described, and the types of inferences they made. I paired this with a bottom-up grounded theory analysis. First, I open-coded observations and inferences in search of themes, then affinity-mapped those themes to form axial codes. Finally, I contextualized those axial codes into a larger selective code. I paired the grounded theory analyses with quantitative analyses of count data.

Participants: 20 participants took part in the study, 10 in the Standard VR condition, and 10 in the Multisensory condition. Participants were recruited through the University of Maryland SONA system, and through word-of-mouth. Participants were 13 women and 7 men that were undergraduate and graduate students, between 18 and 35 years old. Thirteen (13) participants had completed a high school education, 2 participants had completed a bachelor's degree, and 5 participants had completed a masters degree. I chose to recruit from a younger population because I was interested in Millennials as an untapped demographic for citizen science, a contrast to traditionally older citizen science volunteers [9]. I was interested in how amateur water monitors make observations, so I excluded participants with professional experience in ecology.

Experimental setup: During the study, participants made observations and inferences of two stream habitat 360° videos. First, they made observations of a Hot Springs Park in Australia, then they made observations of a California redwood stream. These videos were chosen for their abundance of audio-visual and multisensory phenomena in and around the water habitat, and for their diversity in landscape and climate. As shown in Figure 6.4, the hot spring video features an wide sandy creek with steam rising up from the water. In contrast, the redwood video featured a shady, wadeable stream full of ferns and tall trees.

The 360° videos were 4k high resolution videos collected by Atmosphaeres VR experience website [37], and were imported into the Unity VR player. The Unity setup ran on an Acer Predator i7 gaming laptop in Windows 10. Participants made observations using an Oculus Rift CV1 while sitting and turning their body on a swivel chair. Participants in the Standard VR condition made observations using just the audio-visual headset, while participants in the additional sensory condition made observations using the Holodeck apparatus described in the design section.

To simulate the multisensory habitat of both environments, I researched the audio, thermal, humidity, and olfactory characteristics of both environments, and configured the Holodeck to match each environment's characteristics; the hot spring park has high humidity and high heat, whereas the redwood forest is cooler, with intermittent wind. I did not have access to the exact smells of the hot spring (outback wattle, cow dung, and specific tropical flowers [62]), so in place, I simulated a sweet, humid, greenhouse smell representative of this type of climate. Likewise, I did not have access to redwood essential oils, so I simulated a combination of cedarwood (which grows alongside redwoods) and a sweet mushroom scent, indicative of

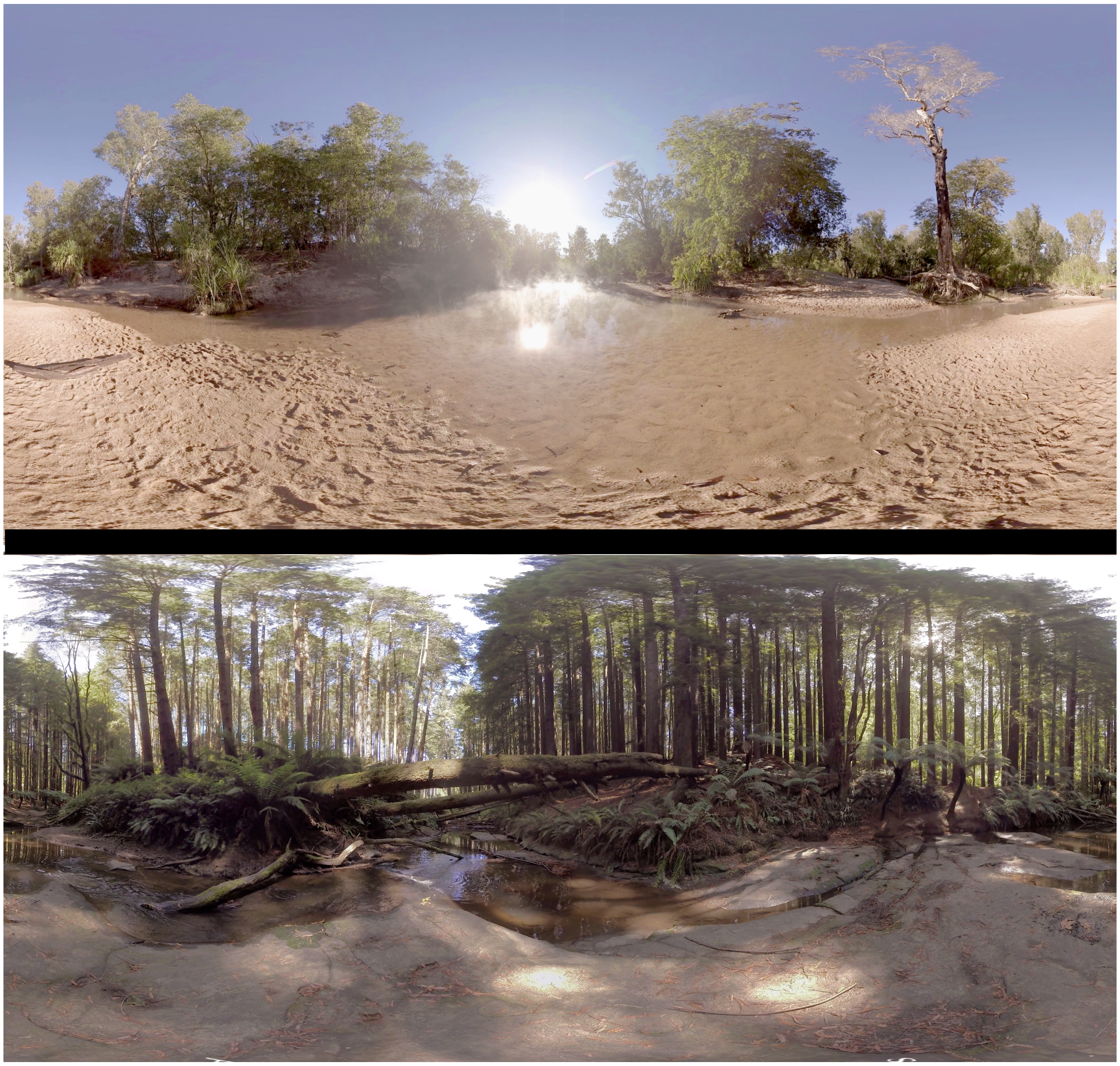

Figure 6.4: Images of the 360° videos I used in the study. Top: Tjuwaliyn (Douglas) Hot Springs Park, Australia. Bottom: California Redwood Forest, USA.

decaying vegetation in an old forest. I also overlaid the audio track with common animal sounds. The hot spring included the sounds of bandicoots, and northern quolls, while the redwood forest included the sound of bald eagles, northern spotted owls, and a mountain lion.

I took care to ensure that the multisensory setup felt organic. To develop a realistic olfactory experience, I pretested different scent combinations, asking users to describe scents with their eyes closed. I also ensured that the combined direction and strength of the sensory information felt natural. I tested combinations of heat, wind, and humidity with the hot spring and redwood forest videos, asking users for feedback on elements that felt out of place.

Procedure: Each study lasted approximately 40 minutes. Participant watched two 360° videos, either in VR, or in VR with ambient sensory cues. With the exception of the Oculus HMD, all sensory components were hidden from viewers behind curtains to prevent participants from being biased by the experimental setup. All participants watched each video 3 times. The first two times, participants made observations aloud about their surroundings. After a short break, participants watched the video a third time, making inferences about the habitat's past and future. As participants spoke, observations and inferences were transcribed word-for-word by a researcher. When necessary, the researcher asked for clarification about words or phrases. After experiencing each video, participants elaborated on their observations and inferences in a series of questions that were also transcribed by the researcher. Participants received a $5 Amazon gift card for participating in the study.

At the beginning of the study, participants were told they were going to watch videos in VR and answer questions about their experience. After consent, participants were told that they were going to make observations of a water habitat in VR, and that they should describe any features of the environment that they saw or physically experienced, and were asked to give as much detail as possible about it. To clarify the types of observations I wanted participants to make, participants were shown a list of 30 examples of visual, auditory, olfactory, thermal, and wind habitat features (e.g. "dam in the water"). The researcher first talked through a few examples in each category with each participants, then participants were given a few minutes to look at this list and ask questions.

For each video, participants first made observations of one half of the 360° environment. They were instructed to look left, right, up, and down without turning their chair, and to make observations of anything they could see, hear, feel, smell, or taste. [1] Then, without taking off their headset, they were instructed to turn their swivel chair 180° and make observations of the second half of the 360° space. I did this to account for literature suggesting that people watching a 360° video do not always inspect the full space [29].

After making observations and taking a short break, participants were instructed to make inferences, not observations, of the habitat. First, I asked participants to make general inferences about the habitat. Then, I asked participants to infer what the habitat may have been like in the past, and how it might change in the future based on observable features. The final time they watched the videos,

[1] Taste was not part of the study, but I included it to prevent participants from focusing on a particular sense.

participants were instructed to turn their chair freely.

After participants made observations and inferences, they were led to another part of the study space, where they were asked to describe the number of senses they experienced, the stimuli they experienced, and what sense was most prominent. Then, on a 7-point Likert scale, they rated how surrounded they felt by the space, how realistic the environment felt, and how familiar the environment felt. Finally, participants were asked to once again describe any inferences they could make about the environment they just experienced. Since participants had limited time to experience the 360° videos, I asked them to make inferences of each habitat a second time, after watching each video, in order to give them time to process and elaborate on their original experience.

After watching both videos, participants were shown images of the two videos (Figure 6.4) with their transcribed observations. Participants first read through their observations, then compared the two environments. First, participants described environment similarities and differences. Then, participants made inferences about the first environment vs. the second environment.

At the end of the study, participants answered questions about their relative enjoyment of both videos, the relative realism of the videos, and their expectation to enjoy future training, and data collection related to stream habitats. Finally, participants filled out questions about their experience with citizen science, water monitoring, and VR. Participants also filled out a self-efficacy survey and answered demographic questions.

6.3.1 Analysis

Study data was analyzed using 1) content analysis of participant observations and inferences, and 2) grounded theory to elicit themes of how participants made sense of the virtual habitats.

Cleaning Data: First, I consolidated and organized observations and inferences made by each participant into 3 categories: observations, inferences, and storytelling. I did this because participants sometimes made inferences while making observations, and likewise made observations when they were asked to make inferences. The third category, "storytelling," allowed me to understand how participants elaborated on observations and inferences through narrative. Each participant's data was thus organized into 3 sections: *Video 1*, *Video 2*, and *Video 1 and 2 Comparison*, and was further organized by observations, inferences, and storytelling.

Observation Level Content Analysis: I was interested in understanding the level of detail with which participants described the stream habitats. To do this, a coauthor and I created a codebook to identify 3 levels of observations. High-level observations (least amount of detail) were coded in orange, mid-level observations (some detail) were coded in yellow, and low-level observations (most detail) were coded in turquoise. In addition, I separately coded observations participants were unsure about in green.

In addition, we counted observations and inferences related to sensory information, and storytelling. We counted the number of sensory—visual, auditory, feeling (humidity, thermal, wind), and smell—observations participants made. We

VIDEO 1

Observations:

- Not too much grass, more single trees or bushes along the banks
- Some birds that can hear but not see
- Still water, not really flowing, Pretty static
- There's fog and steam
- Something coming over from the right
- Trees, looking at creek, but sandy and shallow
- Beached area
- Look like palm trees,
- Kind of like a beachy creek
- Some leaves on the floor, on the sand
- Kind of gurgling in the water
- Feels warm to right side, where sun is supposed be
- There's a streak in the sky
- Now looking in direction of giant tree
- always hard to describe smells, warm humid smell
- Water Is still shallow, ~can see sand and how it goes into puddly area
- Steam is coming from left side, where sun is
- Seems brighter, whiter on left side, bluish on other side
- bird chirping, and animals, little bit of water
- felt warm, but don't know if its my face feeling climate or heat of the machine

Inferences:

- steam is water evaporating because of water and heat
- and the way bank sits, there are some divots, like water was higher
- footprint in the sand, so looks like people were here, with me, that wears shoes, maybe a higher
- Think it's early in the morning, because heating up, but not too warm yet
- didn't see any animals, more of a quiet spot

Storytelling:

- from time to time, people come to this sandbank, hangout because there seems to be footsteps
- future, don't think it would change too much, looks like area, can have heavy rains from time to time, will flood and then return to normal
- the future, either will continue to get dried out, because it looks like some bubbles in the water, evaporating

Figure 6.5: Example data from Video 1, categorized by participant observations, inferences, and storytelling. Observations were coded by level of detail (orange: least detail, yellow: more detail, and turquoise: most detail). I also coded unsure observations in green.

also counted the number of inferences participants made, the number of elaborated storytelling inferences they made, and the number of inferences participants made based on habitat observations.

Figure 6.5 shows a sample of the coded data set, which divides participant data into "high-level," "medium-level," and "low-level" observations. First, we discussed differences in observations, then developed an initial schema for the three levels. Next, we performed an inter-rater reliability analysis on a sample set of 80 observations, reaching an agreement of k = 81.08%, within the kappa statistic norms [84]. Finally, we discussed coding discrepancies and adjusted our coding schema. The coauthor and I each coded half of the data set (5 participants in each condition). During this process, we checked with each other about unclear observations.

Differences in high, mid, and low-level counts were analyzed between condition groups using the Welch two sample t-test, appropriate for small data samples [26]. Since the t-test assumes a normal distribution, I first performed a Shapiro-Wilk test on each of the variables to determine normality. For all non-parametric data, I performed a Wilcoxon-Mann-Whitney test as an alternative to the t-test, which does not assume a normal distribution.

Grounded Theory Analysis: I performed grounded theory analyses on the data to understand emergent trends in how participants made observations and inferences.

1. *Open Coding:* First, I coded the data set with a coauthor looking for emergent trends. During the first round of open coding, we annotated a Google doc with

emergent themes organized by color and condition. After this first round of coding, we discussed these emergent themes. Then, as we were doing line-by-line content analysis, we performed a second round of open coding that more broadly considering our discussion. Coding twice supported further immersion in the data through constant comparison [48]. Open codes included "repetition and elaboration," "inferring one sense from another," and "observations about what they don't see."

2. *Axial Coding:* In order to identify relationships among the open codes, the coauthor and I combined all the open codes into one Google sheet, and grouped them into high-level axial codes using affinity diagram analysis [103]. Axial codes were "immersion," "understanding sensory information," "personal and background knowledge," and "storytelling."

3. *Selective Coding:* Finally, we used the developed axial codes to discern the core variable, "'synthesizing observations," to reflect on how participants made observations and inferences. In doing this, we realized that the axial codes could be redescribed more formally through the core variable, and performed a second affinity diagram process that considered the theory of synthesizing observations to make qualitative judgments. The amended axial codes reconsidered the open codes as part of the qualitative assessment process, "noticing," "making observations," "recognizing patterns," and "forming a cognitive map."

6.4 Results

In this section, I describe my study results. First, I describe thematic outcomes of observations and inferences to understand how participants across conditions made sense of natural watershed habitats. Then, I pair these themes with quantitative results of count and immersion data between subject groups to determine if ambient multisensory information added value to participant observations and inferences.

6.4.1 Thematic Outcomes

My qualitative findings reveal insights about how participants across conditions described the VR habitats. Larger themes were *noticing natural phenomena*, *making observations*, *recognizing patterns*, and *integrating patterns through storytelling*. These themes reflected my core code, *synthesizing observations*, that expressed participants' intrinsic motivation to understand their surroundings.

6.4.1.1 Perceiving Natural Phenomena

Several codes suggested that participants took time to first notice their surroundings. Participants used repetition to describe surroundings, identified unclear environmental information, and inferred multisensory information from visual cues in order to understand the habitat.

I found that participants used repetition to notice and make sense of the virtual spaces, which is in line with previous research about noticing and communicating scientific observations [21, 34]. For instance, participant 10 described they *"[heard] birds chirping, other animals chirping, making sounds"* and said they *"[smelled] something natural, natural smelling."* Likewise, participant 16 described a *"small stream…water is not too fast, has some speed."*

Participants also made inferences about the sensory information from visual cues. For instance, participant 6 said, *"as far as the mist…looks like it's a hot day…could be sunny,"* and participant 7 noted that they *"don't think there is a lot of wind, because leaves are not moving."* Likewise, participant 14 thought that *"it's pretty cool in this area, because [it] looks like it's under the shade."*

Several participants also described features they were unsure about. For instance, participant 20 said that there was *"something in the water, but…don't know what,"* and participant 1 said, *"I don't know how to explain the bird sound, it really sounds like a monkey."* Interestingly, some participants were even unsure about the senses they were experiencing. For instance, participant 5 observed that they *"[smelled] some dampness, [but it] could be in my head."*

6.4.1.2 Making Observations

I found that participants formed more intricate observations as they elaborated on details. These detailed observations often synthesized sensory information, and were made relative to their location in the virtual space.

Participants elaborated on detail through repetition. For instance, participant 10 and 13 expanded their descriptions of the trees as they spoke. Participant 10 observed, *"there's a lot of logs, big logs, with moss, and stems coming out of them,"* and participant 13 noted, *"there's dead trees on my left, huge dead tree."*

Participants made deeper observations by using sensory information, often combining different senses together to form a cohesive observation of their habitat. For instance, participant 16 said that, *"sun is shining through the trees, [I'm] getting some, can feel a little bit of the sun, but not too much,"* and participant 17 thought that, *"steam is coming from my left side, where sun is."* Interestingly, participant 16 elaborated on their observation of a smell in order to understand where it was coming from. They described that a *"muddy smell, like in a forest [was caused by] dead wood."*

I also found that participants made observations relative to their position. For instance, participant 8 said there was *"forest everywhere, ten, twenty feet in front of me"* and participant 1 observed that *"you can see the water running, but it's super shallow because I'm standing in [the] middle of the stream."* Participant 15 added that, *"there weren't people in either [habitat], I was the only person."*

6.4.1.3 Identifying Patterns through Personal Experiences

Participants expanded on habitat observations through patterns, describing these patterns through personal knowledge and experiences. They related observations to locations they had been to, called on their background knowledge to make sense of the habitats, and related memories to understand them.

Several participants often related observations and inferences to locations they were familiar with. For instance, participant 12 said that the habitat *"looks to be like the Amazon, like Jurassic park."* Likewise, participant 15 was *"pretty sure [the] second [habitat] is from a warmer location than Maryland, versus [the] first one [which] could be Maryland."* Similarly, participant 20 suggested that *"this place is probably located somewhere where [the] weather and climate is always hot, probably some place near [the] Amazon river, where there are wild animals [and] a lot of diversity."*

Participants called on their background knowledge to identify patterns. For instance, participant 8 inferred that *"one [habitat is] almost a beach area, very downstream, [because you] don't find...beach on top of [a] mountain,"* and participant 4 predicted the stream in Video 1 would *"probably change shape based on the fact that there are sand bars...[the] river seems to be carving...into the sand, and causing sand to go into the island, and on the other side, being deposited into the bank."* Relatedly, participant 19 used their background knowledge to infer the future of the second stream. They predicted that *"in the future, some company will find out about this area...there's tons of trees, they would use the logs and trees and cut them down. [I know this] not from observation, but prior knowledge, [these are the] types of trees that companies would use."*

Participants also used personal memories to identify patterns. For instance, participant 6 thought that the habitat, *"[looked] like something I would see around Harford County. [I] do a lot of hiking around there."* Likewise, participant 20 said that, *"[I] think this place is humid, because of tiny grass on the branch, [I have seen] this before...not a pretty fancy site, but looks familiar because when [I] go hiking, [I]* *saw a familiar site in Virginia."*

6.4.1.4 Integrating Patterns Into Narratives

Finally, I found that participants integrated observations and patterns into rich narratives. Participants described how habitat details would change over time, even considering how sensory details would change. Several participants told stories to describe the past and future of the two stream habitats. These stories either described natural aging patterns, or the effects of human intervention.

Several participants unified observation and patterns changes into rich narratives. For instance, participant 4 noted that, *"this stream of creek has existed over a long time, creating barrier islands, [while] carving into the mainland."* I even found that participants integrated sensory information into their descriptions of the past and future of the habitats, and considered how sensory details would change over time. For instance, participant 19 thought that they *"would hear less chirping, if more humans would take over,"* and participant 8 inferred that the habitat *"wasn't always a mountain in the past, was a deciduous forest, and turned into a pine forest and mountainous area, and got colder because of the mountain."*

Participants also used stories to describe causation. For instance, participant 2 said that the *"stream was clean water, but either due to pollution, or [a] natural disaster, something happened...[the] water is very dirty and brown."* Similarly, participant 19 described how the forest scene came to look the way it did: *"some of trees branches were cut off...[the trees] didn't have as many branches, so maybe [a]* *lumber company started cutting [them] down, or [they] fell off in a storm."* Likewise, participant 10 considered that *"[the] tree without branches, maybe a fire caused it, [since it] was darker than the others."* When asked about the future, participants were less sure, but used the patterns they'd identified to describe probable scenarios. For instance participant 7 said that *"in the future, water will dry up, because there isn't much water here, or maybe because [this is a] forest...river gets filled up during rainfall."*

Many participant stories described natural aging patterns. For instance, participant 11 said that in the future, *"all the green leaves and trees will become tree branches and stumps,"* and participant 16 thought that *"obviously, [due to the] normal circle of life, [there will be] some dead wood."* Similarly, participant 19 inferred that a *"bunch of fallen logs, that [weren't] there in the past, [they] could have gotten there with lightning, or someone chopped it down to create a path."*

In contrast, some narratives focused entirely on human intervention. For instance participant 2 suggested that *"maybe humans cut the trees down to walk across the stream,"* and participant 9 proposed that *"there are a lot of...logs, so...deforestation nearby, people coming and cutting the trees...would cause [the] water to become muddier, because [there would be] less trees and more soil."*

6.4.2 Quantitative Results

The following section describes my quantitative results. First, I report statistically significant differences between Multisensory and Standard VR conditions on immersion and realism. Then, I report on statistically significant differences in observations and inferences across the two conditions.

6.4.2.1 Immersion

My findings suggested that between conditions, participants in the Multisensory condition were engaged and immersed more than participants in the Standard condition. Multisensory participants felt more surrounded (Mdn = 5.95) by the virtual space than Standard participants (Mdn = 5.15), U = 94.5, p = 7.146e-09. Likewise, participants in the Multisensory condition (Mdn = 5.7) felt that the videos were more realistic than Standard participants (Mdn = 5.1), U = 125.5, p =0.03048.

Participants in the Multisensory condition (Mdn = 5.2) enjoyed watching the videos marginally more than Standard participants (Mdn = 4.6), U = 140, p =0.076. Further, Multisensory participants expected to enjoy future training marginally more (Mdn = 5.1) than Standard condition participants (Mdn = 4.4), U = 136, p =0.076.

6.4.2.2 High, Mid, and Low Level Observations

Overall, I found that the participants in the Multisensory condition made statistically more high-level, mid-level, and low-level observations than those in the Standard condition. In Video, 1, there was a significant difference in the number of high-level observations between Multisensory (M = 10.7, SD = 3.71) and Standard (M = 7.4, SD = 4.09) conditions; t(19.58) = 8.05, p = 1.225e-07, indicating that

Multisensory participants made more high-level observations. In Video 2, participants in the Multisensory condition likewise made more observations; there was a significant difference in numbers of high-level observations between Multisensory (M = 8.9, SD = 3.96) and Standard (M = 6.2, SD = 2.82) conditions; t(9.76) = 7.40, p = 4.091e-07.

I also found that participants in the Multisensory conditions made more mid-level observations than those in the Standard condition. In Video, 1, there was a significant difference in numbers of mid-level observations between Multisensory (M = 19.4, SD = 5.58) and Standard (M = 15.9, SD = 2.28) conditions; t(19.49) = 15.87, p = 1.322e-12, indicating that participants in the Multisensory condition made observations of mid-level than those in the Standard condition. In Video 2, Multisensory participants also made more observations; there was a significant difference in numbers of high-level observations between Multisensory (M = 21.3, SD = 3.52) and Standard (M = 14.5, SD = 3.98) conditions; t(19.39) = 14.43, p = 7.931e-12.

When making low-level observations, I found that there was no significant difference between the number of participant observations in Video 1. However, I found that in Video 2, participants in the Multisensory condition (M = 11.5, SD = 5.95) made more low-level observations than those in the Standard condition (M = 9.5, SD = 4.89); t(19.35) = 7.44, p = 4.322e-07. This suggests that in Video 1, participants in both conditions made similar numbers of low-level observations, but in Video 2, participants in the Multisensory condition made more low-level observations.

6.4.3 Sensory Observations

On average, participants in the Multisensory condition made more distinct sensory observations than participants in the Standard VR condition. Participants in the Multisensory condition made 9.3 observations about feeling (humidity, thermal, wind) compared to Standard VR participants who made on average 1.2 observations. Likewise, Multisensory participants made on average 3.9 observations compared to Standard VR participants who made on average 1.7 observations about smell. [2]

Although including additional stimuli added objectively more features to observe, the mere presence of additional stimuli does not account for the entire difference in number of observations between groups. Participants in both conditions experienced visual and auditory stimuli, however participants in the Multisensory condition made on average 69.2 visual observations compared to 60.2 in the Standard VR condition. Similarly, Multisensory participants made on average 14.2 auditory observations compared to 10.9 in the Standard VR condition.

6.4.3.1 Inferences

In Video 1, participants in the Multisensory condition made more inferences (M = 6.9, SD = 2.73) than those in the Standard condition (M = 5.8, SD = 1.69); t(20.92)= 9.29, p= 7.146e-09. However, in Video 2, I found that there was no difference in the number of inferences participants made.

While there was no significant difference in number of inferences for Video 2,

[2] some participants in the Standard VR condition felt "phantom" stimuli or extrapolated sensory stimuli from audio-visual content.

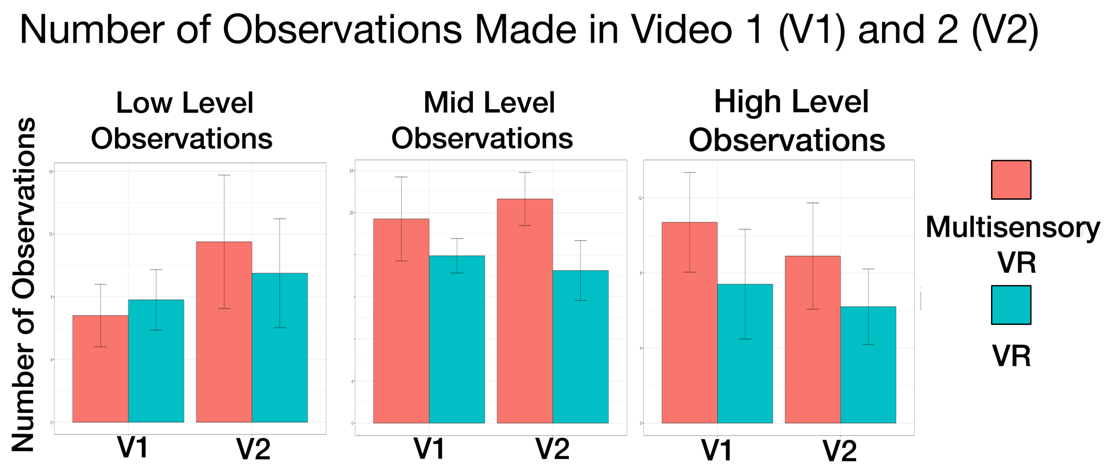

Figure 6.6: Summary of low, medium, and high-level observations. Overall, participants in the Multisensory condition made more observations of their surroundings in the Standard VR condition across in both study videos.

I found that participants in the Multisensory condition made more inferences that were based on habitat observations (M = 5.7, SD = 2.71) than those in the Standard condition (M = 5.2, SD = 2.69); $t(20.43)= 6.55$, $p= 1.963e\text{-}06$.

6.5 Summary

The goal of this work was to consider the role of multisensory realism on the number of observations and inferences that participants made, and on their immersion in the VR stream habitats. My research designed and built the Ambient Holodeck that allowed users to experience environment conditions through olfaction, wind, heat, and humidity. My study then evaluated the effect of multisensory information on participant observations and inferences, pairing bottom-up grounded theory analyses with quantitative metrics.

In my quantitative analyses, I found that participants in the Multisensory condition overwhelmingly made more high-level (least detailed) and mid-level observations across both videos than Standard participants. Further, multisensory information increased the number of observations participants made, and positively affected engagement and immersion.

In complement to this, my grounded theory findings found several important themes; participants across conditions *used narrative to synthesize observations*, first *noticing phenomena*, *recognizing patterns*, and *integrating them* into a cognitive map. My triangulated findings suggest that participants across conditions make observations and inferences using multiple senses; when given more sensory information, participants were likely to use them to better understand habitats.

Chapter 7: Testing StreamBED VR 2.0 against PowerPoint Baseline

7.1 Overview

Through this dissertation, the goal of my work has been to consider whether virtual reality can train citizen scientists to make qualitative assessments of stream habitats. This chapter culminates my design research by testing two versions of the StreamBED VR 2.0 prototype (Standard VR and Multisensory VR) against

PowerPoint training, a baseline similar to standard citizen science protocol training described in chapter 4. In the following sections, I first overview literature on collaborative peer learning, and describe how I designed the StreamBED 2.0 training system. Next, I describe my study method and analyses, and present my study results, pairing qualitative thematic analysis of participant discussion with quantitative analysis of survey responses and participant scores.

The study was multifaceted and complex. Findings between Standard and Multisensory VR conditions reveal that participants in both conditions 1) found the StreamBED VR training system easy to use, 2) enjoyed and benefited from collaborative learning, and 3) were engaged by the system. Comparing all three conditions found that VR participants were more excited for future training, and that Standard VR participants scored closest to a gold standard assessment. Remarkably, the research also found PowerPoint training was also effective when it presented training using the training scaffold I developed.

7.1.1 Collaborative Interaction and Learning

The following section first describes collaborative learning needs, then presents examples of collaborative interaction and learning in HCI. Finally, the section overviews collaborative VR challenges and interaction research needs.

Eberbach [34] contends that experts and novices make observations and systematic comparisons differently; experts notice relevant features, chunk observations, and iteratively ask questions, whereas learners need structure to identify patterns and connect features to function and behavior. In education research, Roschelle [112] asserts that peer-learning between partners with equal backgrounds can help facilitate "negotiation of meaning," a process of constructing meaning incrementally, gradually refining ambiguous or partial meanings through conversational turn-taking that helps learners form constructs of related metaphors.

Collaborative VR has identified different types of collaborative VR spaces. Bouras [10] overviews different types of interactive virtual environments, including Collaborative VE-systems (CVEs) aimed at collaborative tasks, Learning Virtual Environments (LVEs) for collaborative and educational tasks, and immersive virtual environments, such as HMD's and CAVE systems. Relatedly, ShareVR [51] created a platform for asymmetrical interaction between VR and non-VR players, using floor projection and positional tracking. HCI research has begun to consider these collaborative interaction practices in VR interaction, developing many shared design tools and experiences [69, 82, 55].

VR research has likewise begun to design social training in VR. For instance, collaborative VR has been used to teach construction safety [76], train cardiac life support procedures [64], and train pilots to work together while flying [76]. Researchers have also envisioned sandbox VR environments; for instance, CLEV-R [89] mimicked a university setting, allowing college students to interaction with fellow students, and to experience lectures and group projects in VR.

While there is much interest in collaborative learning in VR, collaborative interaction in VR has not yet been standardized and refined. Current psychology and HCI research [50, 49, 51, 59, 121, 70] is in the process of describing the many challenges of VR collaboration; sharing space, coordinating information sharing, and transitioning between individual activities and collaboration.

7.2 Designing StreamBED 2.0

This section describes the iterative design of StreamBED 2.0. First, I outline my design process for three needs identified in previous studies. Then, I describe my system design process.

7.2.1 Design Components

This section is organized by identified training needs: 1) producing environment realism, 2) developing an effective training scaffold, and 3) designing collaborative peer learning. In each section I overview my previous findings, describe my design goals, challenges, and my design process, and outline the final design.

7.2.1.1 Realism

Previous Findings: The first study found that experts made assessments using ambient environmental information, and incorporated multisensory cues into their observations and assessment process. In complement to this, the VR pilot found that participants had trouble making judgments of simulated habitats, and required visual realism to calibrate their assessments. The multisensory study further explored the role of sensory realism in observations, an early part of the assessment process. This study found that multisensory cues helped participants make richer habitat observations.

Goals: I wanted to simulate outdoor habitats realistically. It was important for users to experience VR the way experts experience real habitats.

Challenges: There was a trade-off between interactive and visual realism. Simulating environments produced realistic interaction, but generated low-fidelity visual realism, whereas 360° video created high visual realism, but the 2D videos were not nearly as interactive.

Design Process: To render a visually realistic and interactive VR experience, I tried capturing stream habitats using photogrammetry software, which creates 3D models out of unordered photographs. As shown in the left image in Figure 7.1, I was able to create high quality 3D meshes with a handheld camera. However, when I tried capture a large area using a drone (middle), the software could not create a comprehensive mesh (right).

Final Design: In the final design, participants walked up and down the VR stream by moving through 360° videos connected by hotspots. Participants interacted with the environment by taking snapshots of habitat features.

7.2.1.2 Training Scaffold:

Previous Findings: In the first study, expert interviews and first-person observations revealed that experts had a background in ecology, and formed intuitive judgments of quality by experiencing and comparing features across a variety of stream habitats. In contrast to this, pilot participants had trouble focusing on rel-

Figure 7.1: Trying to simulate realism using photogrammetry techniques. Left: A high-quality 3D mesh of a log I created by capturing several images of it. Middle: A drone I used to try to capture a large stream habitat. Right: A low-quality mesh created from the drone capture.

.

evant features, and required explicit training that helped them separate assessment into actionable goals and tasks.

Goals: I wanted to chunk assessments into clear cut goals and tasks. It was also important for participants to experience a range of stream qualities.

Challenges: Since qualitative assessments are not procedural, a substantial challenge was breaking holistic judgments tasks into piecemeal components. Study training time was also a considerable challenge since participants needed time to experience many habitats.

Design Process: During design, I explored several ways of scaffolding training. In early designs, participants assessed a range of virtual stream habitats guided by non-player characters (NPCs), who acted as expert guides and peer learners. For instance, in Figure 7.2, a NPC character helps learners explore streams, and a crayfish "expert" helps learners calibrate their assessments. These design did not work because of the substantial challenge of creating realistic and dynamic training

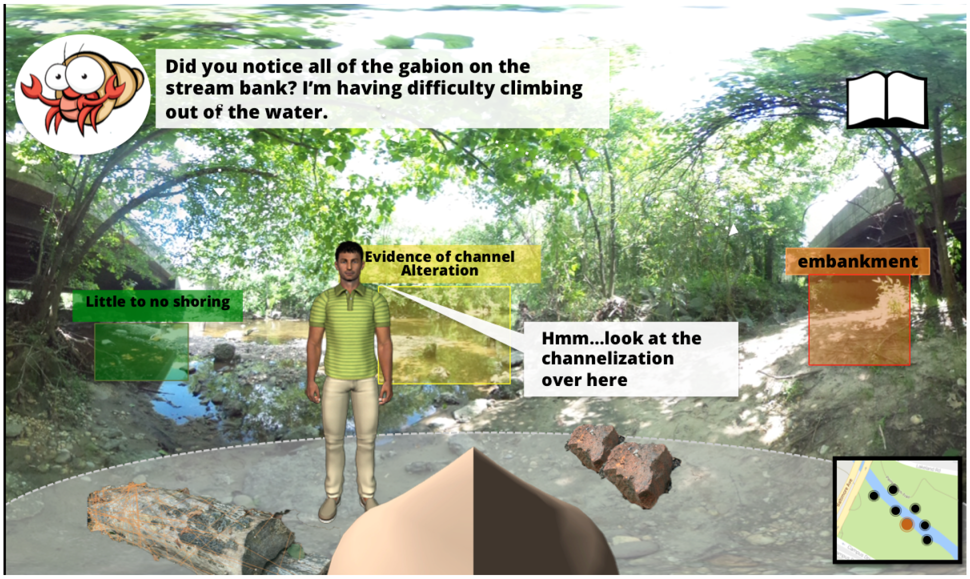

Figure 7.2: An early design of StreamBED 2.0 training. This design featured a crayfish "expert" guide paired with an NPC collaborative learner in a interactive 360° environment.

.

experiences, and because physically exploring several habitats took too much time.

Final Design: The final training divided the assessment task into several small procedural tasks. Participants took snapshots of and tagged important habitat features of a stream in VR, then made a series of feature assessments using reference images on a Desktop application. Participants made a final qualitative judgment by averaging these individual ratings.

7.2.1.3 Collaborative Learning:

Previous Findings: In the first study, interviews found that experts learned to make quality assessments in groups. Further, firsthand observations found that experts made protocol assessments in pairs, using one another to check their observations and assessments.

Goals: I wanted to realistically simulate experts' collaborative assessment experience. It was also important for partners to have equitable experiences.

Challenges: Collaborative training takes more time than individual training. Further, training is subject to scheduling logistics and partner dynamics.

Design Process: My original design considered having both partners interact with training together in VR, but this was overly expensive and hard to coordinate.

Final Design: My final design was an asymmetrical training experience; one participant interacted with VR or Desktop training, while their partner watched interaction on a mirrored display. To give participants equitable experiences, partners switched halfway into each VR and Desktop training task.

7.2.2 StreamBED System Design Process

I designed the StreamBED VR training based on participant needs identified in the previous three studies. At the beginning of my design process, I spent considerable time testing technology affordances to address realism and collaboration needs. I also spent appreciable time considering how to expose participants to a range of habitats in limited time, and how to break the training into standalone chunks.

Once I identified technology and basic training structure, I formed a student development team called *StreamBED Team!* As a group, we went through several stages of design iteration using sketches, bodystorming, and low-fidelity prototyping, shown in Figure 7.3. We tested and refined our designs as we developed the system.

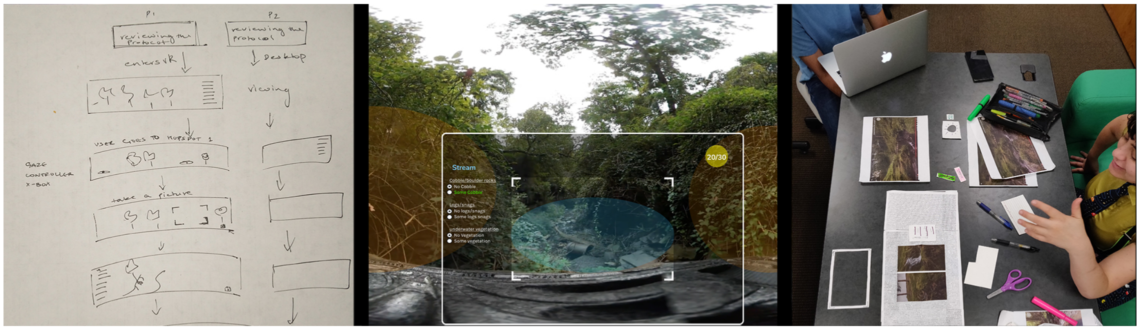

Figure 7.3: An overview of the VR training design process. left: A sketch of the VR training flow. Middle: A low-fidelity VR prototype. Right: Photo of low-fidelity prototyping with the StreamBED development team.

.

7.3 Method

The goal of this study is to consider the value of the final StreamBED VR training compared to traditional PowerPoint (PPT) methods. To consider the value of the final training, I: (1) trained participants who were not expert monitors to make assessments of 2 stream protocols using the StreamBED VR training platform, either with or without ambient sensory cues. I compared training in the StreamBED VR system to a PPT lecture training, typically taught to citizen science monitors [126]. After training, StreamBED VR and PowerPoint participants assessed a real stream on the University of Maryland campus. To ascertain the relative value of the VR and PowerPoint training, I compared outdoor participant scores across Multisensory VR, Standard VR, and PowerPoint conditions to a gold standard score. I contextualized these outdoor assessment scores with engagement and motivation survey responses, and with thematic analyses of survey comments and feedback from VR participants.

7.3.1 Participants:

Thirty-three (33) participants took part in the study, 10 in the Standard VR condition, 10 in the Multisensory VR condition, and 13 in the baseline PPT condition. Participants were recruited using the University of Maryland SONA system and through word-of-mouth. Participants consisted of 19 women and 14 men that were between 18 and 64 years old, however exact ages were not reported. 10 participants had completed a high school education, 13 participants had completed a bachelor's degree, 8 participants had completed a masters degree, and 2 had completed a PhD. Since I was interested only in how non-expert water monitors make observations, only participants who were not expert water quality monitors participated in the study.

7.3.2 StreamBED VR Experimental setup

During the study, two participants worked together to assess a virtual stream habitat through a VR and Desktop training application. First, participants learned about 2 rapid bioassessment (RBP) protocols, Epifaunal Substrate and Bank Stability [6], described in chapter 2. Then, participants completed a VR training tutorial, learning to identify and tag important areas of the stream habitats with appropriate keywords. After completing the tutorial, participants navigated a Maryland stream habitat in VR, taking snapshots and tagging highlighted areas of streams and banks. After the VR task, participants compared the snapshots they had taken to a series of reference images, and made several assessments of the two protocols. After training, participants made evaluations of the protocols at an outdoor stream habitat on the University of Maryland campus.

Figure 7.4 shows the onboarding and training maps and areas. For the tutorial, I used premade 360° videos [54], however, for the training task I needed multiple videos of the same stream. I collected 360° videos of streams using a Samsung Gear 360 Spherical Camera mounted onto a bicycle helmet. The 360° videos were shot in 4k, and were stabilized using Adobe After Effects.

7.3.2.1 Tutorial and Training Videos

Participants navigated several 360° videos during the tutorial and training. First, during the tutorial, participants experienced 4 dissimilar 360° stream habitats around the United states; in Nebraska, New Mexico, Indiana, and Washington State. These tutorial videos allowed participants to experience habitats with diverse stream and bank features. Then, during the training, participants explored 4 videos of a stream habitat in Gaithersburg, MD. These videos allowed participants to "walk" up the stream, similar to how experts explore outdoor habitats. The map and videos for each section are shown in Figure 7.4. Because the study was conducted in the winter, the VR training videos were also captured in the winter.

7.3.2.2 Multisensory Setup

To simulate environments in the multisensory condition, I configured the Ambient Holodeck to match thermal, humidity, and wind characters of the different habitats. I did not simulate ambient smell characteristics of the different environments due to space and time constraints. I took care to ensure that the multisensory setup felt organic, piloting different ambient conditions with the videos.

The first tutorial environment (top left image in Appendix 7.4), a Nebraska stream, was cool with a morning breeze. I simulated this environment with two fans pointed at the VR participant from different directions. In contrast, the New Mexico stream environment was warm and humid. I replicated this environment with two heat lamps pointed at the VR participant from different directions paired with a cool mister. I reproduced the third tutorial stream, a shady area in Indiana, with one heat lamp pointed at participants' back, the mister, and one fan. Finally, in the warm, lush, and windy Washington stream, I used one heat lamp, the mister, and both fans. In contrast to the tutorial streams, the Maryland stream was wet and cold, so I paired the two fans at a high setting with the cool mister.

7.3.3 Procedure

Each study lasted approximately 2 hours and 15 minutes. During the study, participants learned about the two protocols, completed a VR tutorial, and completed a VR training where they photographed and tagged important stream areas. Then, in a Desktop application, participants assessed the tagged snapshots based on protocol features, and made an overall evaluation of each protocol based on the individual assessments. Finally, participants evaluated an outdoor stream on the University of Maryland campus. Study procedures are summarized in table 7.1.

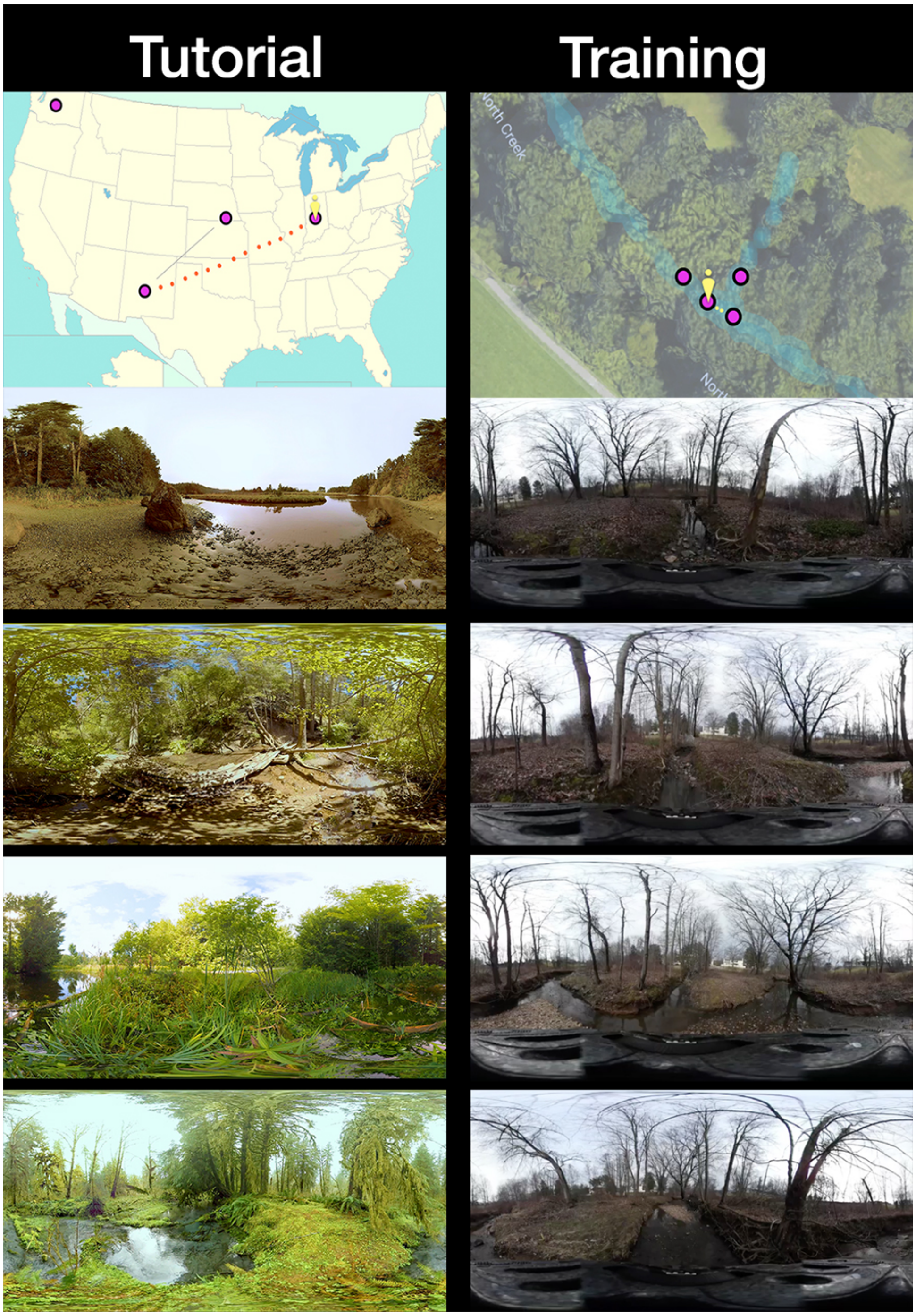

Figure 7.4: Screenshots of onboarding and training maps and areas. Left: Screenshot of onboarding map and four 360° stream locations participants experienced; in Nebraska, New Mexico, Indiana, and Washington State. Right: Screenshot of map and 4 locations of a stream habitat in Gaithersburg, MD.

.

Participants answered survey and discussion questions about their experience after training and evaluation. Participant comments and questions were also audio-recorded during the study. When necessary, I asked for clarification about words or phrases. Participants received a $5 Amazon gift card for participating in the study.

Table 7.1: Final Study Procedures

Time	Study Description
5 Minutes	Study Introduction and Consent
20 Minutes	Offline Introduction to Assessment Protocols
20 Minutes	VR Tutorials
30 Minutes	VR Stream Assessment
30 Minutes	Desktop Feature Assessment
30 Minutes	Outdoor Assessment

7.3.3.1 Offline Introduction to Assessment Protocols

At the beginning of the study, participants were introduced to one another, and were given an introduction to citizen science, water quality monitoring, and two protocols, Epifaunal Substrate, and Bank Stability [6]. Participants were told they would be learning to make qualitative stream assessments because local watersheds fed directly into drinking water, and it was important for local community members to understand how stream habitats were affected by human intervention.

Then, participants were shown a notebook with illustrations of the two protocols, paired with the actual protocol scale, shown in Figure 7.8. These illustrations were created to clearly show protocol key terms. Participants listened to a description of each protocol and keywords, then were asked to discuss and mark up the illustrations and scales, and to ask initial questions about the protocol. Participants were also told that they would learn more about assessment as they went through training, and were not required to memorize the protocols.

7.3.3.2 VR Tutorial

After familiarizing themselves with the two protocols, participants completed a VR tutorial, then completed the VR training together. During the tutorial, one participant experienced the training in VR, while the second participant watched their mirrored display on a large monitor. This setup is shown in Figure 7.5. Although the VR participant controlled the experience, their partner supported their experience by commenting on what they were seeing, and giving feedback on what areas to photograph and tag.

I created this asymmetrical co-learning relationship based on co-learning literature that suggests that having two learners verbally discuss a problem made it easier for them to understand [112]. Both participants experienced the role of the VR participant and partner, switching halfway through the tutorial and training.

Before beginning the tutorial, participants received an overview of the Oculus controller and VR interactions. Then, the first VR participant put on Oculus setup, and opened the first tutorial scene from the "Menu" screen. Once the VR participant opened the first scene, they saw a map of the tutorial locations (Figure 7.4). Then, the VR participant closed the map to see the "immersive view" of

Figure 7.5: Left: In-game view of Oculus Controller. Right: VR participant and partner working together to tag the Maryland habitat in VR.

.

the 360° stream habitat, and looked around the stream with their partner. After looking around, participants opened their camera, and searched for areas highlighted areas in the camera viewfinder (Figure 7.6). These highlighted areas were either blue (representing stream areas), or yellow (representing bank areas).

Once participants found the highlighted areas in the viewfinder, they took a "snapshot" of the area, zooming in and out to center the area in the viewfinder.[1] After taking the snapshot, a set of keywords appeared on the left side of the viewfinder, stream keywords if the highlighted area was blue, and bank keywords if the area was yellow (Figure 7.6).

After taking a good snapshot of the highlighted area, participants worked together to decide which keywords best represented the highlighted area in the snapshot. For blue stream areas, participants had to decide if the area included any 1) snags and logs, 2) cobble, or 3) underwater vegetation. Similarly, for yellow

[1] Participants could discard any snapshots not centered on the area they wanted to capture.

Figure 7.6: The viewfinder and tags in the tutorial and training. Left Top: The viewfinder over a blue "stream" area. Left Bottom: the snapshot of the stream area with stream tags (logs and snags, cobble, and underwater vegetation) that appear after the snapshot is taken. Right Top: the viewfinder over a yellow "bank area." In contrast to the tutorial, multiple bank areas appear next to one another in the viewfinder. Right Bottom: A snapshot of the bank area tagged with "some undercut banks," "some crumbling," and "some exposed tree roots."

bank areas, participants had to decide if 1) the bank slope was relatively gentle or steep, 2) the bank was vegetated by grass, plants or trees, and if the bank had any 3) undercuts, 4) crumbling, or 5) exposed tree roots (Figure 7.6). As participants discussed each area, they were given supporting instructions to notice habitat features and choose appropriate tags.

As well as making participants comfortable with the VR controls, the goal of the tutorial was to make sure participants understood what the different keyword tags represented. Together, participants discussed the tags, made a choice for each keyword category, and saved the tags. In the tutorial, participants received feedback on the tags they had chosen; if they correctly chose tags, the snapshot was saved. However, if participants incorrectly tagged the area, the incorrect keywords turned red, and they were prompted to tag the snapshot again.

Once participants correctly tagged the highlighted area and saved the snapshot, the highlighted area glowed white (instead of blue or yellow), indicating it had been saved. When participants tagged all of the highlighted areas in each 360° video, they moved to the next video by aiming at a hotspot on the screen. At the end of the last video, participants were directed to a summary screen that allowed them to see how many times they had tagged each keyword.

7.3.3.3 VR Training:

During the VR training, participants explored a Maryland stream through a series of four (4) 360° videos. The training mirrored the tutorial interactions; upon entering a new scene, the VR participant saw where they were on the map, and closed it to experience the full 360° immersive view of the stream. Then, they opened the camera, found and took snapshots of highlighted areas, and tagged them. Unlike the tutorial, which featured one highlighted area per video, the training included many highlighted stream and bank areas to capture and tag.

As in the tutorial, the VR participant and their partner worked closely to look around the space, choosing areas to take snapshots of, and choosing appropriate tags for each snapshot. The two participants had different viewpoints—some participants found it easier to see details from the mirrored monitor, while others found it easier to see the areas in 360° —so each relied on the other for perspective. Similar to the tutorial, participants switched halfway into the training, so each had a chance to experience both partner roles.

At the end of the training, participants saw a summary screen of their tags (Figure 7.7) organized by protocol keywords, rather than by "stream" and "bank," [2] and made an initial assessment of the Epifaunal Substrate and Bank Stability protocols.

The VR participant first summarized the tags for the first protocol, and the Desktop participant read the protocols aloud to the VR participant to remind them of the scales. Next, participants used the tag summary and protocol descriptions to make an initial rating based on the relative number of features participants observed.

[2] Even though they were organized differently, the two protocols closely matched the "stream" and "bank" keyword organization. Epifaunal Substrate included all the stream keywords with the additional of "undercut banks," and Bank Stability included all the bank keywords with the exception of "undercut banks." I chose to slightly modify the keywords to stream and bank to make it easier to tag the areas.

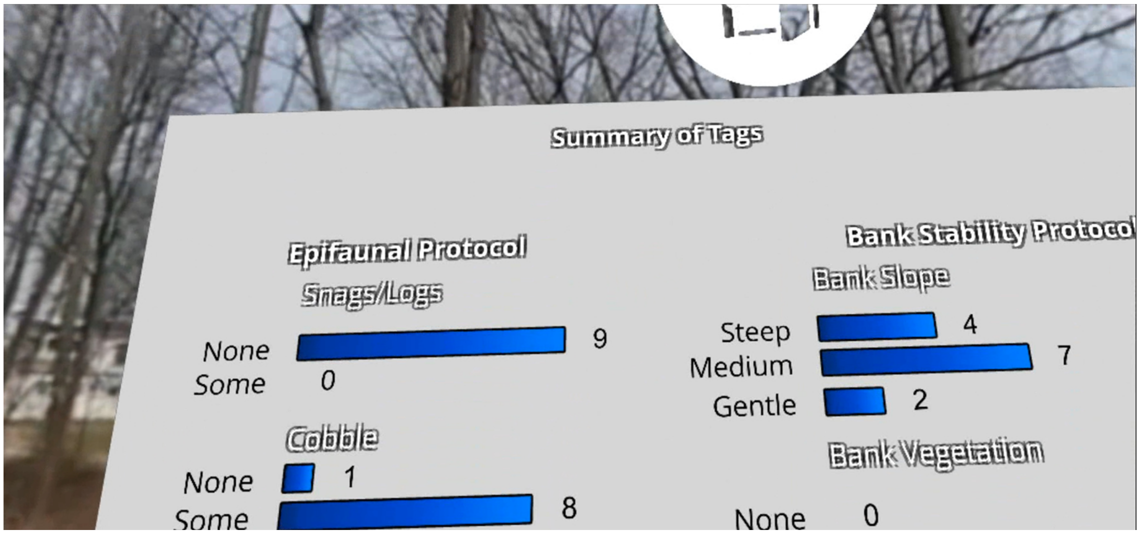

Figure 7.7: VR training summary screen. The summary screen showed the number of each tags participants made in each category.

.

Participants were instructed to first discuss what protocol category ("Optimal," "Suboptimal," "Marginal," and "Poor") best represented the number of features they observed in the stream habitat, and then to narrow down their choice to a single number. [3] The non-VR partner marked this initial assessment in blue marker on the physical scale (Figure 7.8). After rating both protocols, the VR participant exited training by pointing to a hotspot above the summary board.

At the end of the VR training, participants completed a short survey of their experience as the VR participant and as the partner. As the VR participant, participants answered questions about environment immersion, environment realism, and their ability to interact with their partner during training. Likewise, as the partner, participants answered questions about their engagement and ability to contribute to the VR task.

[3]Both scales use percentages in their protocol metric descriptions. This initial assessment was meant to be a representation of quantity (but not quality) of each of the areas.

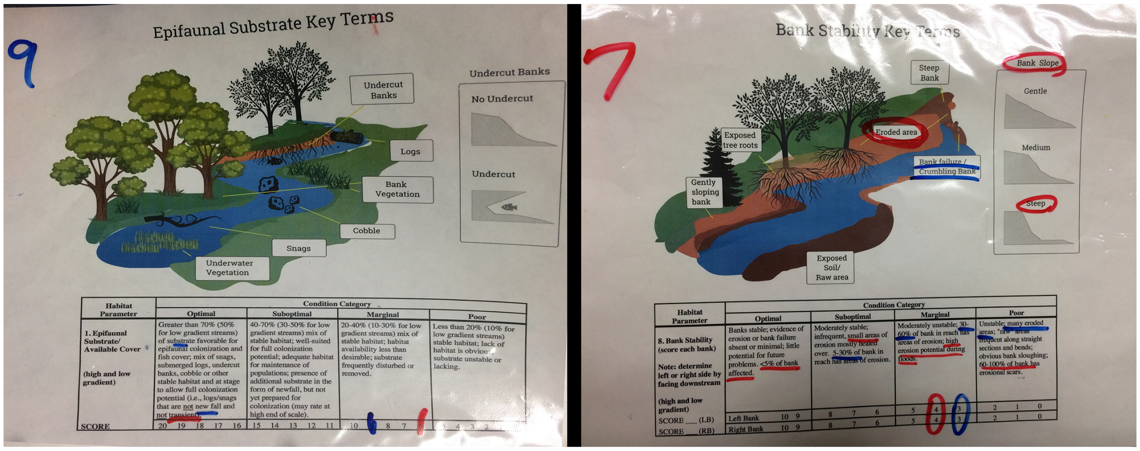

Figure 7.8: VR training protocols for Epifaunal Substrate and Bank Stability measures. Participants marked up both protocols during the introduction, and then used the pages to make the initial assessments (in blue) and final assessments (in red). Left: Epifaunal Substrate protocol assessment. Right: Bank Stability protocol assessment.

7.3.3.4 Desktop Assessment:

Participants compared and rated the snapshots they took in VR using reference images. First participants selected a subset of images that represented each feature, then compared and rated each Epifaunal Substrate and Bank Stability feature based on reference images. Participants used these individual feature assessments to make a final assessment of each protocol. Similar to the VR tutorial and training, participants switched between being the “Desktop” participant making decisions, and the partner, who participated through a mirrored display.

During VR training, participants experienced and tagged the same areas of the stream and bank from different perspectives. During the Desktop task, participants first selected the images that best represented each feature in each area. For instance, in Figure 7.9, Participants had to choose which of the two images from area 11 best

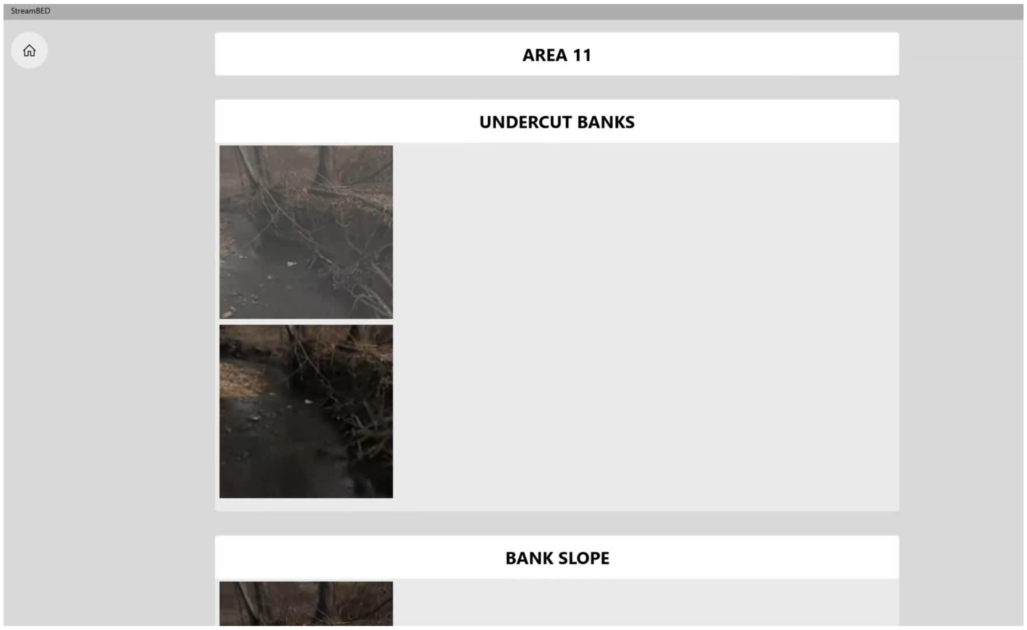

Figure 7.9: Screenshot of participant selecting the best image that represented the feature "Undercut Banks" in Area 11.

represented the feature "undercut banks."

Once participants selected a representative image for each area, they compared the snapshot features to a set of reference images, and rated them on their associated protocol scale. For instance, the feature "cobble" is part of the Epifaunal Substrate scale, which is rated on a 20-point scale. For this feature, participants compared their snapshot to 5 references images of cobble, showing what cobble looked like at *20, 15, 10, 5,* and *0.*

Figure 7.10 illustrates this interaction. During this feature assessment, participants used the left and right arrows to scroll between the five reference images. Each reference image was surrounded by a border showing the general area to focus on. When participants scrolled over the reference image, the evaluated feature, "Cobble," was highlighted in the image, and paired with a description of subfeatures that contributed to its rating. In Figure 7.10, the reference image has a rating of *15* because it has relatively more cobble in the stream, more vegetation and moss on the cobble, and because many cobble are partially submerged in water.

Participants were encouraged to consider the reference descriptions alongside the reference images. They were told that the reference images were meant to be a good relative representation of the numerical scale, however, that one image could not fully capture all of the subfeatures that contributed to a given rating. For instance, a 15 on the scale could have a lot of cobble that was submerged in the water, but not covered in vegetation. Likewise, a 15 could have fewer cobble that were densely covered in vegetation and partially submerged.

Before they could make an assessment for each feature, participants were required to first scroll through all of the reference images. [4] After scrolling through the reference images and looking through the scrollover descriptions, participants looked for a reference point to make their assessment. If participants felt an image was close to a reference point, they could rate the snapshot with one of the reference numbers. However, if participants believed an image fell between two reference points, they clicked the "+" sign to choose a rating in between two reference points. Participants made feature assessments sequentially; they evaluated all the snapshots tagged with one feature before rating the next feature.

After evaluating Epifaunal Substrate features, participants evaluated Bank

[4] This was done to ensure that participants actively used the reference images and descriptions to make assessments. The "submit" button was greyed out before participants scrolled through all of the reference images, to prevent them from accidentally make an assessment without comparing their snapshot to the reference images. The submit button was only greyed out for the first snapshot in each feature. After navigating through the reference features a first time, participants make other assessments of the same feature without scrolling through reference features.

Figure 7.10: Screenshots of participants rating the "cobble" feature. In both screenshots, the snapshot being evaluated is on the left, and a reference image and description for a 15-point rating is on the right. Top: the snapshot and reference image side by side, with an outline of the stream area. Bottom: When participants scroll over the reference image, the reference image highlights the cobble in the stream, and shows a description of the subfeatures that contribute to the rating.

.

Stability features. Interaction for the second protocol was identical to the first protocol with the exception that Bank Stability is measured on a scale of 10 to 0, instead of 20 to 0 [6]. For this protocol, participants saw 4 references, representing *10, 6, 4* and *0.* [5]

Making the Final Assessment: Participants made a final assessment of each protocol by pairing their individual feature assessments with their initial assessment. Shown in Figure 7.11, participants saw a final assessment screen for the Epifaunal Substrate protocol, displaying the feature assessments they had made as icons on a number line. [6] The figure shows that participants made 5 "Cobble" assessments and 7 "Undercut Bank" assessments. The figure also shows 16 assessments for the features "Snag and Logs" and "Underwater Vegetation"(not pictured) at 0. In the Epifaunal Substrate protocol, a feature not being present, and therefore not tagged, was equitable to the "worst" quality feature. These untagged features were automatically represented as a "0" on the final scale.

As the Desktop participant navigated the final assessment scale, their partner was given the physical protocol scale marked with their initial assessment. At this point, participants were prompted to make a final assessment based on 1) their initial assessment and 2) the range of individual feature assessments. Participants were advised to equally weigh the different features, and to consider how the quantity and quality of the judgments they had made compared to the scale description.

[5] Unlike the 20-point Epifaunal Substrate scale, the 10-point Bank Stability scale did not evenly divide, making it difficult to map the reference images to the scale. I chose to add more reference images at 6 and 4 in order to add granularity to the center of the scale.

[6] Participants could access a key showing the icon-to-feature mappings from a key icon at the bottom right of the screen. Additionally, participants could look at the Epifaunal Substrate protocol by clicking on a paper icon.

Once both participants agreed on a numeric assessment, the partner marked the number in red on the physical protocol (Figure 7.8), and the Desktop participant chose the same number on the Desktop application. After submitting their final answer, participants received feedback about how close their final score was to an expert's score (Figure 7.11). Participants then completed the same steps for the Bank Stability Protocol.

After Desktop training, participants filled out a survey about their experience. After rating their experience with the Desktop training, participants answered overall questions about the VR and Desktop training.

Outdoor Evaluation: After the VR and Desktop training, participants walked to the Paint Branch stream on the University of Maryland campus to complete the outdoor evaluation. As participants walked to the stream, they answered open-ended questions about their training experience with their partner. These open-ended questions are provided in Appendix A.5.

Figure 7.12 shows two participants exploring the Paint Branch stream as they make assessments. At the stream, participants were each handed a marker and a laminated copy of two protocols, shown in Figure 7.13. Similar to the training, participants were told to work together to complete the outdoor evaluation, but should make their final assessment independently. The study was conducted in February, so participants were told to do the best they could to account for the winter season. [7] Participants were told they could ask for help, but completed this

[7]Expert water monitors traditionally conduct RBP assessments in the summer, but logistics prevented the study from being conducted in that season.

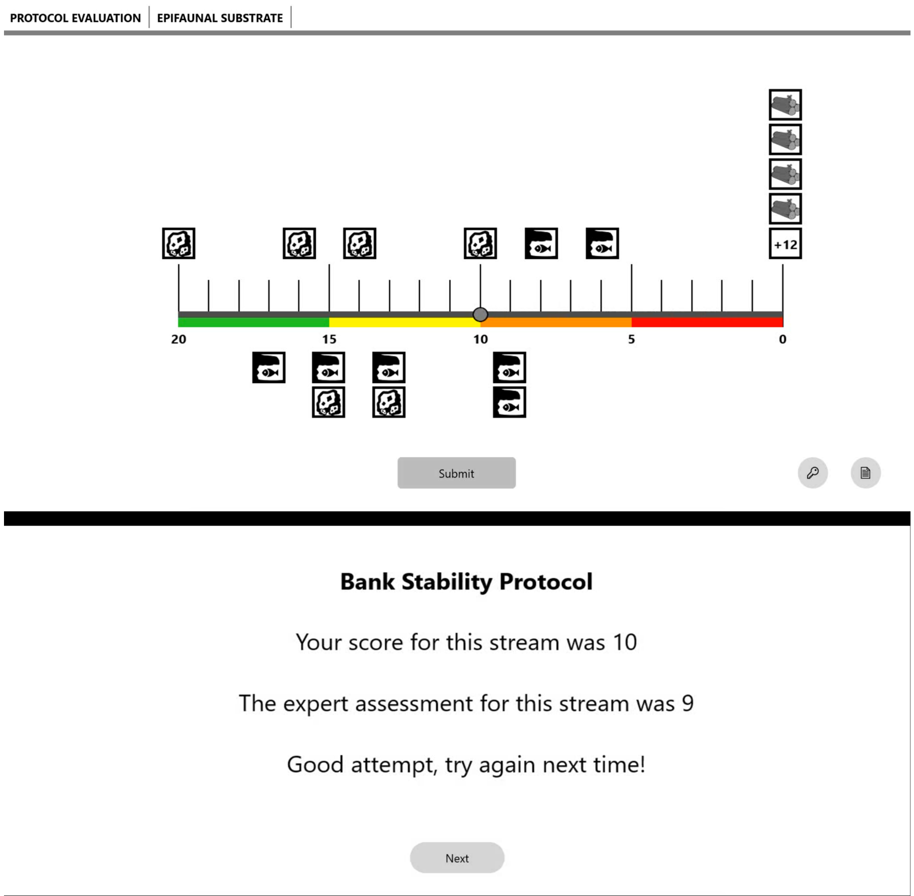

Figure 7.11: Final assessment and feedback screen for the Desktop application. Top: Final assessment screen for Epifaunal Substrate. The number line shows the individual assessments participants made. Participants made a final assessment by dragging the gray circle along the number line. Bottom: Feedback screen for the Bank Stability rating. Participants received feedback on how their final score compared to an expert's.

Figure 7.12: Two study participants evaluating the Paint Branch stream at the University of Maryland.

.

final evaluation without external guidance.

After completing the outdoor evaluation, participants walked back to the study space. As participants walked from the stream, they answered followup questions about their training and outdoor experience. These open-ended questions are also in Appendix A.5. Once back in the study space, participants answered questions about their outdoor data collection experience, background questions about their citizen science and water monitoring experience, and demographic questions.

7.3.4 Baseline PowerPoint Experimental Setup

The baseline PowerPoint study was a control condition that reflected the way citizen science volunteers are currently taught about the RBP protocol [126]. A more detailed description of this firsthand experience is described in Chapter 4. The goal of the baseline study was to match the traditional PowerPoint format using VR study content. During the study, a class of participants first learned

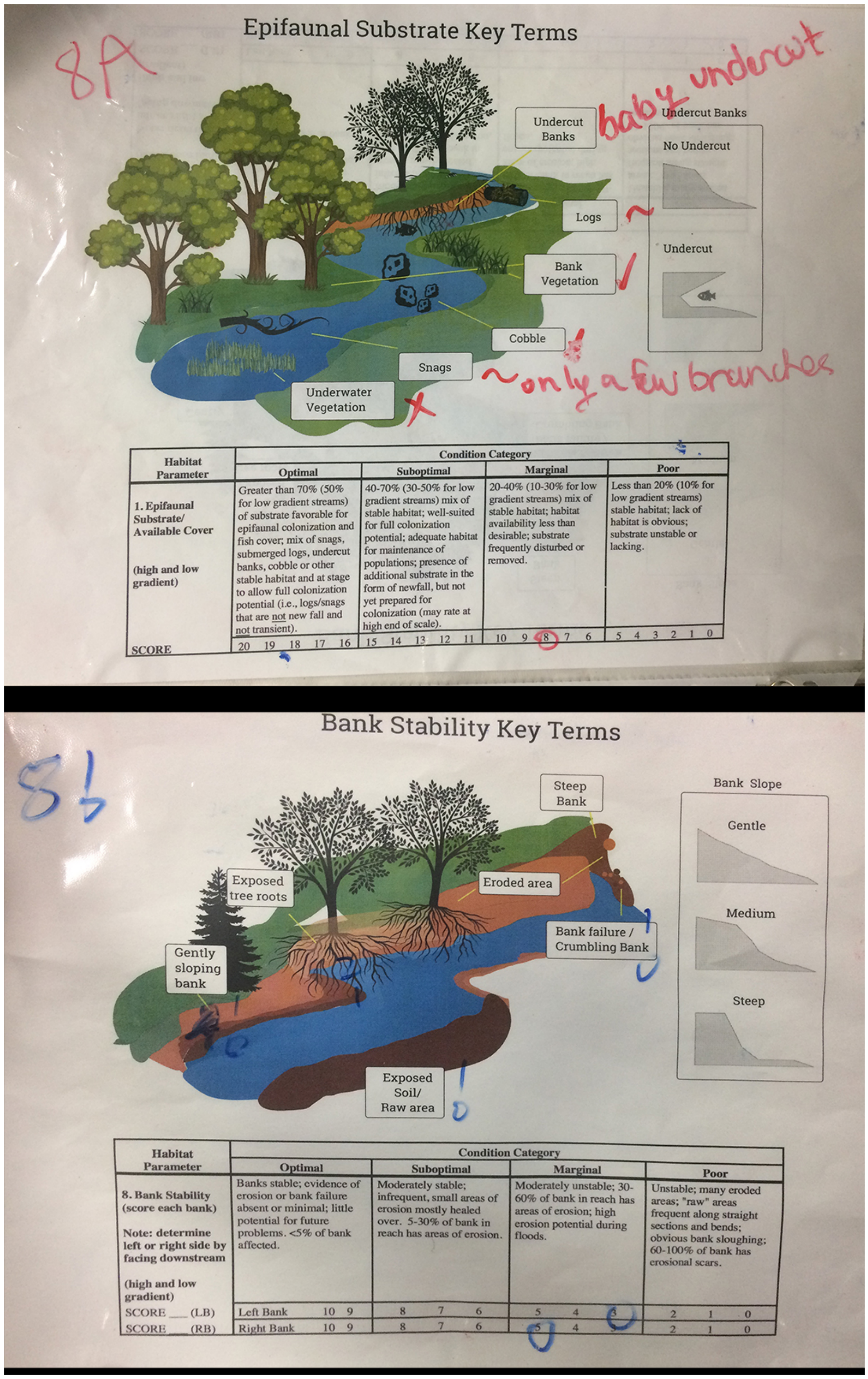

Figure 7.13: Examples of individual outdoor assessments completed by VR participants. Top: Epifaunal Substrate assessment completed by participant 8A. Bottom: Bank Stability assessment completed by participant 8B.

.

about the Epifaunal Substrate and Bank Stability protocols, made assessments of the Gaithersburg stream through a series of images, and made an assessment of an outdoor stream at the University of Maryland.

PowerPoint Training: Similar to the VR study, PowerPoint participants were first given an introduction to citizen science, water quality monitoring, and two protocols, Epifaunal Substrate, and Bank Stability [6]. Similar to the StreamBED training, PPT participants were told they would be learning to make qualitative stream assessments because local watersheds fed directly into drinking water, and that it was important for local community members to understand how stream habitats were affected by human intervention.

Then, participants were introduced to Epifaunal Substrate, the first of the two protocols. Participants listened to a description the protocol and keywords, and then asked initial questions about the protocol and keywords. Participants were told they would learn more about the measures as they went through training, and that they were not required to memorize the protocols.

After participants learned about the protocols and keywords, they saw a set of reference images showing how each keyword mapped onto the protocol scale (Figure 7.14). These images mirror the reference images StreamBED study participants experienced in the Desktop application. Participants first saw all of the reference images for snags and logs, undercut banks, cobble, and finally stream vegetation. After seeing the reference images for each protocol, participants had a chance ask questions. After learning about the Epifaunal Substrate protocol, participants likewise learned about the Bank Stability protocol. First, they learned about the protocol

Figure 7.14: Examples of the PowerPoint study training images. These images mirror the VR training reference images.

.

and keywords, and saw reference images for bank vegetation, bank crumbling, bank slope, and exposed tree roots.

After learning about both protocols, participants identified protocol keywords in a series of images. Participants were shown screenshots of the same areas StreamBED participants experienced during the VR tutorial, and were asked to call out whether they saw those features. After learning to identify the stream and bank features, participants saw screenshots of the same Gaithersburg stream that VR participants experienced, and as a group, made Epifaunal Substrate and Bank Stability ratings of the same areas that VR participants experienced. At the end of the study, participants made a final evaluation of both protocols as a group, and received feedback on how close their final score was to the expert score.

Outdoor Evaluation: After the PowerPoint training, participants walked to the Paint Branch stream to complete the outdoor evaluation. Participants were instructed to make evaluations of the stream in pairs, as shown in Figure 7.15. At the stream, participants were each handed a packet with the two protocols (similar

Figure 7.15: Pairs of PowerPoint training participants assessing the Paint Branch stream on the University of Maryland campus.

.

to Figure 7.13) and a pen. Similar to the VR conditions, PPT participants were told to work in pairs to evaluate the outdoor habitat. Like VR participants, PPT were asked to make their final ratings independently.

After completing the outdoor evaluation, participants filled out survey questions with pen and paper (see Appendix A). Similar to the VR study, they answered questions about their outdoor data collection experience, background questions about their citizen science and water monitoring experience, and demographic questions.

7.3.5 Analysis

Study data was analyzed using qualitative thematic analysis of VR participant outdoor discussions, and quantitative analyses comparing the two StreamBED VR conditions, and comparing the VR conditions to the PPT baseline.

7.3.5.1 Thematic Analysis

Thematic analysis was performed on participant discussion with the researcher that took place while walking to and from the outdoor stream (see Appendix A.5 for discussion questions). Discussion data was transcribed and broken up by participant utterance.

The goal of the thematic analysis was to uncover emergent trends in the two StreamBED VR conditions. First, I created a initial codebook, then worked with a second coder to 1) code data using preexisting codes, 2) identify and apply open-codes, 3) develop study themes using affinity diagrams, and 4) diagram code relationships. Due to the complex nature of this research, coding resembled an ethnographic process that included lengthy discussion and checking with the second coder.

First, I created an initial codebook of codes based on previously identified themes. I did this because of the study's complexity, and because the second coder had not been previously involved in the project. Shown in Appendix A, this codebook consisted of 17 codes grouped by *realism*, *co-learning*, *scaffolding learning*, and *overall training interaction*. Each code included a set of related questions, and an example of the code in a positive and negative valence.

Before coding, the second coder read through through the codes and had an opportunity to ask questions. Then, we coded a sample of of the dataset and performed an inter-rater reliability analysis (K=80%). We amended the codebook description and examples based on discussion after this process. Next, we coded the full dataset using the amended codebook. As we coded the full data, we identified bottom-up open codes that did not fit the preexisting codes. Having identified these bottom-up codes, we did a second pass through the full data set identifying these open codes in the data.

After coding the initial themes and bottom up codes, we combined all the open codes into one Google sheet, and grouped them into high-level themes using affinity diagram analysis [103]. High-level themes were *Training Scaffold / Framework*, *Immersion*, *Personal vs. Collaborative Decision Experience*, and *Process Developing Expertise*. While the themes fit the open codes well, we felt that more complex relationships existed between the four themes. For this reason, we created and iterated on a diagram (Figure 7.16) showing the web of relationships between the themes.

7.3.5.2 Quantitative Analyses

The goal of this research was to understand how different training experiences affected participant engagement, interaction, and learning. I used quantitative analyses to compare participant engagement, interaction, and scores across the three conditions. First, I described participant engagement and interaction in the three conditions, then I compared common metrics. After describing the training experiences, I compared participant outdoor assessment experiences in the three conditions, describing differences in engagement, interaction, and participant scores.

Due to the small sample size (approximately n=10 per condition), my quantitative analyses were limited to descriptive statistics and analysis of variance. First, I report participant means for the different variables, and compare differences between groups. Since ANOVA assume a normal distribution, I first performed a Shapiro-Wilk test on each of the variables to determine normality. For all non-parametric data, I performed a Kruskal-Wallis Test, a non-parametric alternative to ANOVA that is appropriate for small data samples [1]. Differences between participant and the "gold standard" score was calculated by taking the absolute value of the participants' score subtracted from the expert score.

The StreamBED VR engagement and PPT training were measured on different Likert scales due to the different distribution methods. The VR metrics were measured on a 7-Point Likert scale, however some of the PPT metrics were measured on a 5-Point Likert scale.[8] For this reason, I report the two metrics differently; when reporting differences between the VR conditions, I report them on the original 7-Point scale, and when comparing the 5-point VR and PPT metrics, I convert the 7-point scale to a 5-point scale. Table 7.2 shows the corresponding 7 and 5-point Likert scores. 7 was scaled to 5, 6 and 5 were scaled to a 4, 4 was scaled to a 3, 3 was scaled to a 2, and 1 remained the same. In the quantitative results I clarity when I use the 7 or 5-point Likert scales.

[8]I ran the PPT study with paper questionnaires rather than online using Qualtrics. I chose to limit the number of variables that PPT participants responded to because on paper, it was difficult to present all of the questions on a 7-Point Likert scale in a way that easily fit onto paper. Further, without showing participants a visual scale, it was difficult to anchor participant responses to text (e.g. "not at all," "somewhat," and "very much.")

Table 7.2: Rescaling From the 7-Point to 5-Point Likert Scale

	7-Point Likert Score	5-Point Likert Score
Least	1	1
	2 and 3	2
Somewhat	4	3
	5 and 6	4
Most	7	5

7.4 Results

In this section, I describe my study results. First, I describe thematic outcomes of observations and inferences to understand how participants in the standard and multisensory conditions experienced the training and outdoor assessments. I pair these themes with quantitative results of engagement and assessment data to consider whether there was a difference between the Standard VR, Multisensory VR, and PowerPoint conditions. I summarize my findings after each section.

7.4.1 Thematic Findings

7.4.1.1 Thematic Outcomes of Observation and Inferences

My qualitative findings reveal insights about how participants in the standard and multisensory conditions experience the training and outdoor assessments. Through content analysis and open coding, I identified 26 individual codes, and integrated them into larger themes using affinity diagrams. Larger themes included *Training Framework*, *Immersion*, *Personal vs. Collaborative Decision Experience*,

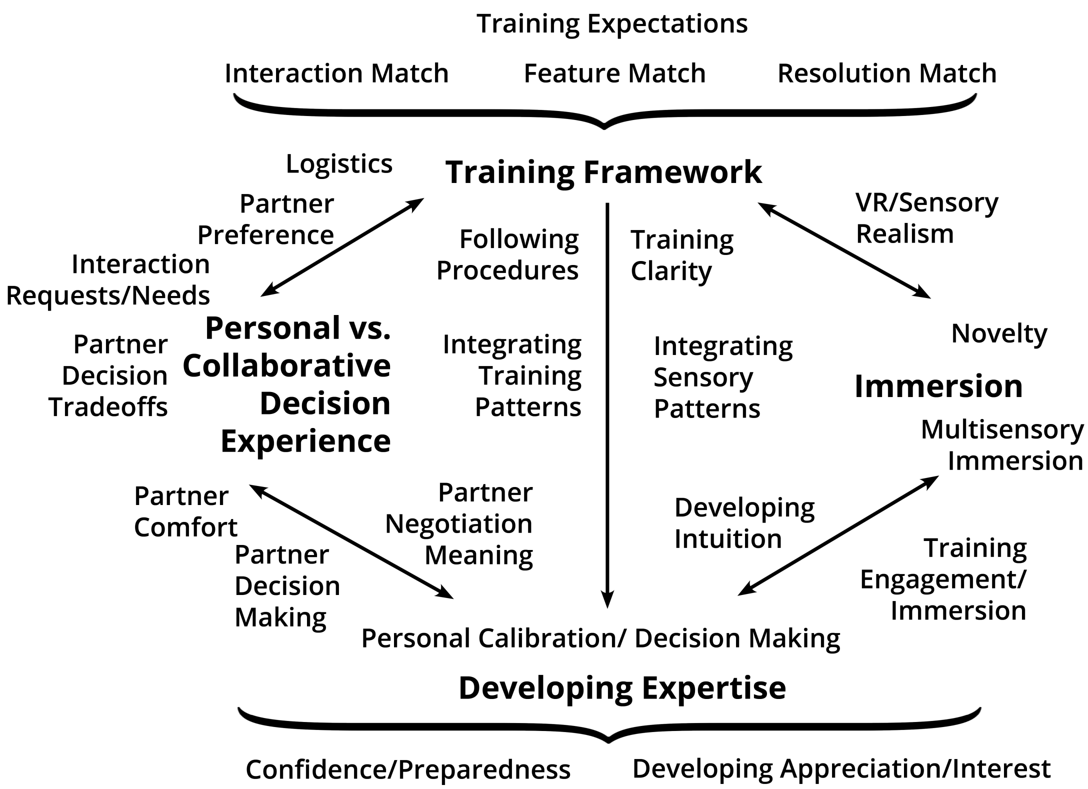

Figure 7.16: Thematic diagram showing the complex relationships between the four themes.

.

and *Process Developing Expertise*. Together, these themes provide insight into how participants interacted with training, and how this affected their ability to develop expertise. During affinity diagramming, a coder and I found that the relationships between the different themes were closely connected. To show these relationships, we diagrammed the themes and individual code relationships. In the next section, I first present the complete diagram of relationships, and describes its organization. Then, I reconstruct the diagram as I discuss each theme.

7.4.1.2 Connecting Themes

Our coding process found that the themes were connected through a network of complex relationships, shown in the thematic diagram in Figure 7.16. This diagram shows the relationship between the primary themes, *Training Framework*, *Immersion*, *Personal vs. Collaborative Decision Experience*, and *Process Developing Expertise*, and the individual codes that make up each theme.

The themes and codes are organized by stage in the learning process and by relationship proximity. The *Training Framework* theme describes the training experience, affecting participants' *Immersion* and *Collaborative Experience*. Together, the training framework, immersion and collaborative experience affect participants' process of *Developing Expertise*. Codes directly above the Training Framework theme (Expectation, Interaction Match, Feature Match, and Resolution Match) represent inputs into participant experiences; participants come in with an expectation of training difficulty, and interact with training components. Likewise, codes directly under Developing Expertise are outputs; the process creates internal feels of confidence and preparedness, as well as an appreciation and interest for the material and the participants' surroundings. Similarly, codes between the Training Framework and Developing Expertise theme describe the stages of Developing Expertise from top to bottom; participants first learn to follow procedures based on training clarity, integrate training and sensory patterns, negotiate meaning, and build intuition. Finally, participants make decisions, with their partner and by themselves.

The codes are also organized by relationship proximity; codes close to one

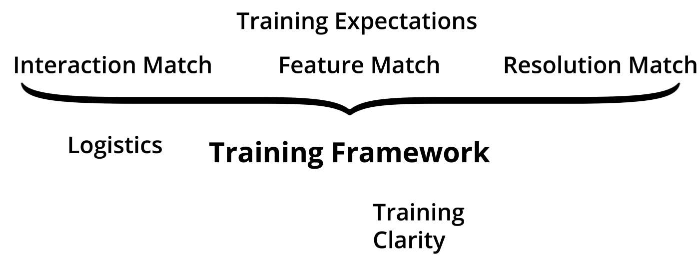

Figure 7.17: Thematic diagram showing the Training Framework codes.

.

theme are primarily connected to that theme, whereas codes between multiple themes are connected to all of those themes. Study logistics affected participants' preference for working with a partner, and participants' comfort with their partner affected how learners negotiated meaning and made decisions together. Likewise, training realism and novelty affected how participants integrated sensory patterns, and how engaged and immersed they were. The summation of these experiences contributed to participants' personal decision making, through which they developed expertise.

7.4.1.3 Training Framework

During the study, participants were asked to reflect on their experience with training, first directly after training, then after data collection. When asked about their experience, participants described the StreamBED training as clear and easy to follow, and considered whether training interaction, features, and resolution matched the outdoor assessment. Participants also made requests for future training, and considered logistics of the co-learning and VR experience. The codes for this theme are presented in Figure 7.17.

Overall, participants described the training as clear and easy to follow. After the training, and before the outdoor assessment, participant [3B] remarked that "*for the heuristic, actually for that I had no idea had to do that...when you gave [the training] to me...it doesn't take much effort to get it*" and [9A] and [9B] agreed that "*the UI was so easy to navigate...if you made a mistake, you almost couldn't make a mistake.*" [6A] was particularly appreciative of the tutorial, explaining that "*I have not used a remote since I was 14...I don't play video games...it was nice to have that moment to like get adjusted visually.*"

Later, having experienced the outdoor assessment, participants likewise remarked on the clarity of the training. Participant [6B] was "*surprised at how easy it is to pick it up. I really thought it was going to be more difficult to like pick up the VR and...assessment.*" Interestingly, [7A] and [7B] remarked that the training was more effective than a course they were taking that semester. "*we take...landscape architecture...this semester...the information that was similar...and the knowledge we learned today was much better...a lot more clear...I learned a lot more than I learned in the lecture,*" explaining it was because of the combination of "*the VR, and also, the pictures on the computer...and like, the real thing.*"

Interaction, Feature, and Resolution Match

After the outdoor assessment, participants relayed mixed reactions to how well training interaction, features, and resolution matched the real outdoor experience. Some participants thought the training experience reflected the outdoor experience well, while others commented on differences in interaction. For instance, [9A] said, "*I think in general [the training] did do the job of introducing us to different elements to pay attention to,*" but noted that the training interaction was more detailed than the outdoor assessment. "*The training, the places we had tagged were segmented...when we did the part of the assessment*" they explained. *"But when we were outside...I didn't segment it at all...specific places...how it was in the training.*" Likewise, [7B] thought the training was "*a lot more detailed...we talked a lot more than we did [outside].*" Participants also noted how the differences in training and outdoor interaction affected their ability to make judgments. [8A] explained that "*I had a hard time breaking it down areas [outside], because...the left bank had a gentle slope, but the right side, if you looked right, on the left bank...it was more steep and more eroded...so I was like, I'm having to super generalize this whole left bank.*"

Some participants noted the positive benefits of the difference in the training experience. Participant [5A] considered that the VR experience "*provided the opportunity to go get a different angle...l [looked] at it holistically down the length of it...see how it was changing,*" whereas outside "*...we were confined to one, one side of it [the bank].*" Likewise, [6A] and [6B] thought that training made it easier to experience different parts of the stream. They felt that " *we were able to see areas of the stream that we wouldn't have been able to see in an in person assessment...even walking up and down the banks.*" Similarly, [8A] noted that during the outdoor assessment, their view "*[was] obstructed,*" explaining that "*there was no way [to see to right of a stream bend]...unless we wanted to get wet...[whereas]...In VR...I don't feel that any [chance] of like...falling in the water.*"

Study participants also recounted how differences in resolution and features impacted training. Participants [8A] and [8B] said that compared to training, the outdoor assessment "*definitely has a better resolution...seeing [the stream] in person...is like high rez.*" Participants also noted that certain features were easier to see in the real world than in VR. For instance, [9A] and [9B] described how shape and contour of the banks were easier to see outdoors, and [3A] observed that murky water in the training was challenging to observe. "*I think...It's harder to know what's going on down...underwater*" because of tree reflection in the water. "*If I were actually...in the reality, then I would have used some object...to kick the dirty water.*"

Participants also considered how well environment features matched between the training and outdoor assessment. Participants thought the Maryland stream videos were similar to outdoor assessment, but had trouble comparing the snapshots of the Maryland stream to reference images in the Desktop task. [3A] said that "*its [hard] to account for the seasonality of the environmental conditions...I think most of the photos are taken right now, right? But the reference photos are mostly taken from summer.*" Similarly, [9B] said that they thought "*the pictures were really good pictures, but I don't think [the seasons] matched up.*"

Considering Study Needs and Logistics

During the study, participants identified several training needs and discussed the logistics of the VR experience. When discussing training needs, participants wanted a deeper way to coordinate and engage with one another during training, and wanted more detailed reference points.

Participants in groups 8 and 9 wanted to be "*be able to point at things [in training]*" [9B], specifying that "*it was hard to explain...what part of the picture I was seeing*" [8A]. Participant [9B] elaborated on what they wanted by describing a collaborative game mechanic. "*You know how when you play video games, if you keep going off track...[there's a pointer] that kind of guides you toward...the other player.*" Surprisingly, some study participants creatively improvised on the training interaction, using "*the corner of the picture...panel*"[8B] to point to environment features. Relatedly, participants in Group 3 wanted to be able to show one another what part of snapshots they were focusing on, wanting to "*be able to zoom the photos after you took them.*"

In Group 5, participants requested more detailed descriptions of reference images. Participant [5A] explained that "*I don't necessarily know what technically qualifies as undercut...or...exposed...if I were doing this more, I would want the actual technical definition, of these are the boxes to check...to differentiate what's esoteric to that image as opposed to what's indicative of the feature*"[5A]. Participants further explained that they wanted "*multiple [reference] pictures...rather than one picture*"[5B]. The group further explained that they wanted to see a cluster of images at each reference point. [5A] explained that it was important "*being able to differentiate what's esoteric to that specific image as opposed to, oh, all of 20's have this in common...maybe a third of them also have this [feature] in common.*" As [5A] described this need, [5B] suggested that the training could use machine learning to extrapolate some of these patterns. For instance, an algorithm could show "*what* *[features] showed up more...if there was something common in both.*" Related to Group five's request for detailed images, Group 9 wanted more detailed feedback at the end of the training. [9A] wanted the "*the...end...the expert score...[to be] more complicated*" so the group could better understand how far they were from the expert score, "*why we scored it that way.*"

As well as describing their needs, some groups considered the logistics of the VR and collaborative training. Some participants felt that they could do the training without VR, while others felt that they could have done the training independently. In Group 10, [10A] thought the training could have been taught without VR. "*I don't think this is easier than just using plain screen,*" they explained. "*You need to be at a particular space to use VR...it's special equipment. If I...can use just the website, then I can...work from everywhere.*" Group 10 further elaborated that reasoning was practical; "*If I have easy access to the VR system, then yeah, why not?...if I can afford it, then I would...be perfectly fine with it.*"

Groups 8 and 10 also weighed the value of collaborative learning against scheduling costs. When asked whether having a partner was valuable during training, [8A] responded that "*I think it definitely helps to get someone else's feedback, but I think I could do it on my own.*" Participant [8B] explained that "*if I do it voluntarily, in my...free time I need to arrange my time and schedule everything...I'm gonna do it on my own.*" When asked about working with a partner, [10A] and [10B] acknowledged that working with a partner is "*better, but there's a coordination cost...especially for graduate students, it's really hard to find a time that works for everyone.*"

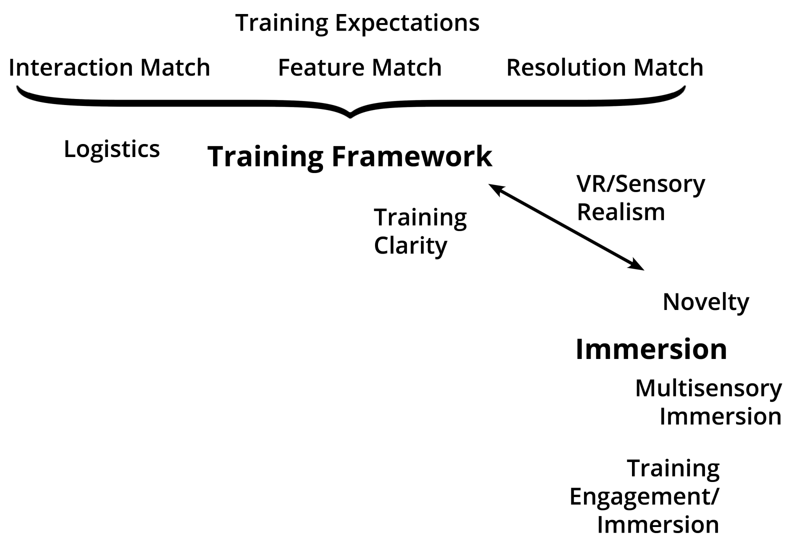

Figure 7.18: Thematic diagram showing the relationships between Training Framework and Immersion Codes.

.

7.4.1.4 Immersion

As they discussed their training experience, study participants described their engagement and immersion in the training. Participants reacted to the realism of the VR experience, and commented on the novelty of the training. Multisensory participants also described the sensory immersion of additional multisensory stimuli. The codes for this second theme, and its connection to the Training Framework theme, are presented in Figure 7.18.

VR Realism and Novelty

Several participants felt the training was realistic and immersive. Participant [7A] responded that the training "*feels like you're literally in the video...surrounded by...the environment is easy to assess...it was realistic.*" Participant [3A], who had experienced the pilot StreamBED training, was "*surprised by the quality of the VR* *training of the streams...cause I took part in your previous study...and that was a huge improvement...both [parts of the training were] immersive.*" Notably, [10B], who initially questioned the logistics of VR training, changed their mind about the value of training immersion and realism after completing the outdoor evaluation. "*When [I'm]...doing all of these things through VR stuff...it's like...I'm experiencing everything,*" they explained. "*I can move my head around and see what is actually going on and what is actually in front of me...So it's much more like a real experience than just looking at the web browser or something like that.*"

Several participants reacted to the novelty of VR and multisensory stimuli. For instance, [8A] thought "*it was cool to be able to turn in 360 and look at everything in either environment,*" and [8B] was excited by the tutorial's "*tropical...like environment...this is so cool.*" Likewise, [6B] said the multisensory immersion was"*really cool...so also, when you put the heat on, and like the wind and everything...it made me feel like I was there.*" In contrast, [5B], who was used to VR, commented on how their unique experience contrasted other participants. "*I have a VR arcade back where I work...So I'm kind of used to VR,*" reasoning that "*for some people...it takes time to [get used to] the VR...they're very overwhelmed with everything...for me, it's just like, okay I'm looking at these photos.*"

Multisensory Immersion

When asked about the additional sensory stimuli, multisensory participants [5, 6, 7, 8, 9] were split on how they perceived the sensory stimuli, and whether it affected their training immersion. Some participants did not notice or feel affected by the multisensory experience. For instance, [5A] responded that "*I'm not neurotypical...I'm very verbally focused...and so when we were doing this study and I had the VR headset on, I literally stopped being cold...I couldn't feel the steam and I couldn't feel the breeze...because that part of my brain stopped processing that information because it wasn't important.*" Participants [8A] and [8B] said that they could feel the heat stimuli, but did not realize it was a part of the study, explaining "*maybe it's like hot in the room...I was thinking...that you're trying to cool it down. But I did, I remember thinking...the light is like giving me a tan.*"

In contrast, some participants found value in the sensory stimuli. Participants [9A] and [9B] said that they "*enjoyed [the multisensory stimuli]...I thought it was interesting...[and] made me excited about the training.*" Participants [6A] and [6B] were even further immersed by the sensory stimuli, recalling that the sensory information made them pay attention to the training. [6A] explained that "*it felt you were in the environment...I know that it was the point of VR...but sometimes I feel like when I look at the screen...I kind of glance at it...but this...I was taking more time to take it all in.*" Likewise, [6B] commented that without "*feeling the nature...in VR...I feel like it would have seemed less real to me...and more like a game,*" adding that "*I probably wouldn't have...thought about it [the training] as hard.*" Together, the two participants summarized that the sensory stimuli "*made [the training] feel more important.*"

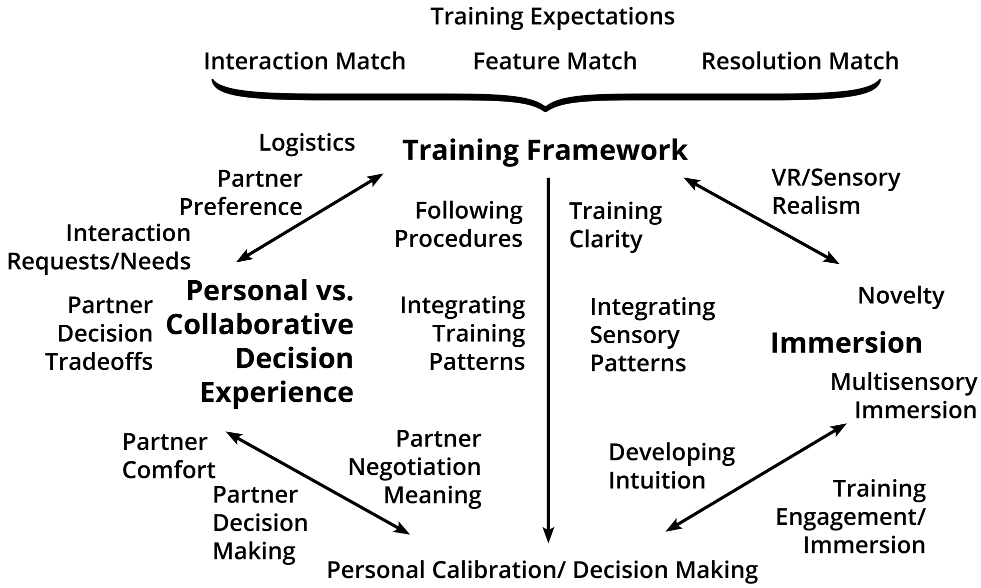

Figure 7.19: Thematic diagram showing relationships between Training Framework, Immersion, and Collaborative Experience codes.

.

7.4.1.5 Personal vs. Collaborative Decision Experience

During the study, participants were asked to reflect about their collaborative learning experience. Participants described their partner experience in relation to their decision making process; how comfortable they were with their partner affected how they negotiated meaning during the training and outdoor assessments, which in turn affected their personal calibration and partner decision making. In this complex experience, participants described a wide range of partner experience preferences, learning and making assessments with the same partner, alternating partners, or working independently. The codes for the Collaborative Learning theme and its' connection to the Training Framework and Immersion themes are presented in Figure 7.19.

Partner Comfort

Participant pairs had a range of relationships; some signed up for the study as partners or friends, while others were paired with an unfamiliar person. After training, participants were asked to describe if their partner familiarity affected their training dynamic, and whether they preferred working with someone they knew. Learners described a range of comfort levels; [2A] and [2B], who were friends, responded that having a partner they knew "*felt much more comfortable...[we were] not afraid to say something wrong.*" Similarly, [7A] and [7B], who also knew one another, agreed that "*I think if we work with...a partner we don't know...[inaudible]...we don't know...how to communicate with them.*"

In contrast, some pairs that knew each other preferred to work with partners they did not know. [10B] felt that "*because we know each other...I feel a little bit uncomfortable to...to have some directly opposite kind of opinion.*" Their partner, however, was unconcerned about the dynamic, responding, "*I don't have a good opinion...even If I know somebody or not...I always oppose [laugh].*" Relatedly, pairs that did not know each other felt comfortable working together. For instance, [8A] reflected, "*I don't think it was a bad thing that we didn't know each other. I didn't feel like we had to, to complete the tasks.*"

Negotiation of Meaning

A primary task of co-learning was to make sense of the training and assessment environments together, reflecting how experts monitors learn to evaluate stream habitats together. Co-learning partners navigated the training and assessment habitats together, describing the value of having a partner during the process, and revealing differences in training and outdoor assessment dynamics.

During training and assessment, participants appreciated having a second person to help them "*know what [features] you're looking for*"[3B]. For instance, [7B] and their partner "*talked about whether there is...underwater...vegetation...and whether there is logs.*" [10A] clarified that this negotiation process helped them calibrate their "*understanding about the criteria. Like when I was thinking...about the vegetation...on the banks...[my partner] was telling me like...oh for the first evaluation...you have to focus on the...Epifaunal thing.*" Participant [6B] added that "*it was nice to have someone bounce off of.*"

In VR, participants had some trouble coordinating this process with one another. For instance [9B] explained that "*whenever I wanted to observe something...it was hard for me to point*" and [9A] added that "*where...I was looking...I could only give her directions...left...left*" As described in the training requests section above, several participants (8A, 8B, 9A, and 9B) requested a pointer to help them navigate training. Outside, however, participants had no trouble coordinating these actions.

Calibration and Decision Making

A step after negotiation of meaning, participants described how they made assessments decisions. Participants also considered how this experience was affected by their partner, and discussed partner decision trade-offs.

During the assessment, participants calibrated their assessment based on expert rating feedback. For [3A], the training "*was more about learning what's expected* *for this rating, more about calibrating my evaluation criteria.*" They elaborated that based on the expert training feedback, "*I was pretty generous at first...evaluating the epifaunal substrate...but then after training, I realized that [I was]...getting closer to the reality...to the accuracy.*" Similarly, [8B] described their group's use of the reference assessment; "*I was think about that reference...like the difference between our scale and the reference...that we did this morning.*"

Participants also tried to calibrate their assessments to the seasonal differences between the reference images (taken in the summer) and the assessment tasks (in the winter). Because of this difference, some participants felt that the "*the reference frame was too good to be true*"[3A]. [8A] and [8B] described this difference; "*[the reference images were] almost not comparable*" to the training and assessment environments, they explained. "*Some of the high rating [reference] pictures...look like wallpaper.*" Participants responded that if the references images matched the season, their assessment experience would be easier.

Participants said that having a partner pressured them to make thoughtful judgments, and helped them calibrate their decisions. Participant [3A] explained that "*timewise, I would be much faster if I did it on my own, but I think quality wise it's better to have second eyes to cross check if there's something I missed.*" Similarly, [8B] felt that without a partner, "*I might [make assessments]...in a lazy way...lose focus...spending too much time on each figure...[and] overlook some features of the training test.*" They added that "*having someone as a partner, kind of pressed to me to focus a little bit more...because it's not wasting my [own time].*" Participants also thought that having a partner would help them calibrate their judgments. Participant [7A] said that having a partner made them make "*better decisions...because [if] you're not sure...can ask the other partner,*" and [3A] added having a partner helped calibrate any "*difference in your answer.*"

Although having a partner could improve decision making, Group 10 felt that the partner dynamic negatively impacted their decision making process. [10A] explained that "*in terms of calibration...having a second person is good...[but] if one person is dominant...in opinions...then it's not good.*" The pair explained that, "*when we were [rating] VR things...I felt like...it's more like...one person tried to persuade another.*" For this reason, when they did the outdoor assessment, the participants chose not share their assessment scores, even though they were knew they could. "*It happened naturally...we were just talking about...the facts...not persuading each other.*" Strikingly, the pair remarked that this dynamic only occurred because they made independent ratings, and acknowledged that "*if we need only one rating...it would be...more precise.*"

Partner Preference

Finally, participants gave a range of feedback on whether they preferred collaborative or independent future training; some preferred working with their partner, others wanted to switch partners, and some participants wanted to try independent training.

Several participant groups [3, 5, 6, 7, 8, 9] preferred working with a partner during the training, explaining that it "*would be more fun*"[4A], and that "*it was helpful to get someone else's opinion*"[8A] and "*it was easier to make judgments*"[9A] because "*looking at different...environments...I might be kind of biased*"[4B]. When asked whether they preferred working with the same partner, participant opinions split. Some preferred the partner they were comfortable with, while others appreciated the opinion of a new partner. For instance, [6A] and [6B] agreed that "*if you're working just with one person and you get increasingly comfortable with them, you guys can start to think the same way. But if you change partners, you'll get...a different perspective that you haven't thought of before.*"

Other participants felt that they were ready to try independent training. Although [4A] said that working with a partner was enjoyable, they felt that "*I could probably do it on my own.*" Likewise, [3A] relayed that "*I feel more confident that I could be better at it...this time and attempt it on my own.*" Some partners wanted a combination of independent and assessments. "*When we make decisions we can...[be] influenced by our partner's decision,*" [7A] explained. "*I want to make sure...I understand the material first, and then I can check [the assessment]...with the partner.*"

7.4.1.6 Process of Developing Expertise

The Training Framework, Immersion, and Collaborative Learning experiences blended into participants' development into an expert water monitor. Participants began training with an expectation of the training, and first learned how to follow the procedural training tasks. As participants followed this procedure, they assimilated environment patterns, and constructed a sense of intuition. This intuition created a sense of confidence and preparedness, contributed to task enjoyment, and increased their overall interest in the world around them. The codes for the final theme are presented with the other three themes in Figure 7.16.

Training Expectations

Several training participants began the study with expectations about task and training difficulty. For instance, [2A] was "*initially…kind of nervous…oh God, I'm going to have to like, learning stuff, especially about…ecology, earth science, and stuff like that,*" [6B] was "*stressed about [the training],*" and [8A] felt "*so overwhelmed when you were explaining the VR…I was just intimidated when you start talking about it, I was like, wow, I don't think I'm going to be able to follow all these steps.*"

Overall, participants were surprised at how easy and intuitive the training material was. Participant [2A] was relieved that the training was "*a lot easier than I expected…it was good…once I got in there, it was super intuitive, so I really liked that.*" Participant [6A] was likewise "*surprised at how easy it is to pick it up…I really thought it was going to be more difficult to…pick up the VR and do the assessment through VR…once you did it once or twice, I feel like it became more intuitive.*" Participants [7A], [10A] and [10B] similarly commented that the training was easier than they expected.

Following Procedures

As they went through training, participants became comfortable with the procedural tasks. [6A] explained that they "*got comfortable with everything…the terminology, and the technology.*" Participants then explained how they internalized the assessment process. "*When you're doing the outdoor assessment, what I noticed is that we know where to start the assessment...we should start from the erosion, we can start from the undercut...it sort of gave us a guidance*" [3A] explained. [8A] further elaborated that "*there were some keywords on the paper...and I [went] over them sequentially,*" then looked "*at the assessment criteria...the evaluation table...and then...at the status in the stream...[and finally] made the marks of scores on each [tier].*"

Having gotten comfortable with the protocol, several participants expressed that future training would be faster. Participant [3A] said, "*I feel like I'm fairly well trained now...I think a second time should take less time, much less,*" [4B] thought future training "*would be slightly faster,*" and [6A] and [6B] agreed that "*it definitely would be faster for me...it would be faster...practice makes perfect.*"

Integrating Patterns

The next step in participants' process of developing expertise was integrating visual and sensory patterns. When integrating visual information, participants knew "*what [features] I am...looking for*" [3B], but were not entirely sure how to integrate and interpret environment information. [3B] explained said that "*it was harder to...elaborate...what the logs looked like [8A]*" due to seasonal differences in training features. Some participants also had trouble judging the Epifaunal Substrate and Bank Stability protocols because of unrelated features. For instance, [4B] was confused by artificial bank shoring (that was part of a different protocol) at the assessment stream. "*I wasn't prepared for the...rocks along the other bank,*" they explained.

Participants also needed more practice recognizing environment differences between habitats, and integrating them into a total score. Group 2 acknowledged that "*there's still a lot that we're very unsure...like with the undercuts, we thought that they were very exposed,*" but felt that making assessments "*would definitely become easier over time...once we practice more with recognizing the difference between the environments.*" Likewise, Group 5 knew what they were "*looking for, but not necessarily how to interpret what I'm seeing correctly yet.*"

The sensory information helped some multisensory participants to remember and transfer details between training and assessment. For example, [7A] said that the sensory stimuli helped them remember the training; "*when you go outside, and the wind and the heat...kind of makes you think of the training...it's also easier to feel like...it's the same kind of environment,*" explaining that the sensory cues helped them "*remember, and recall*" the training. Relatedly, [9A] focused on how temperature affected visual information; "*sensory wise for me, I noticed that it was cold...and how it would affect the surroundings...more like [how]...moss...looked like when it was cold,*" and [9B] noted that just "*knowing about the temperature helped me justify something...things like, there's no vegetation.*" Notably, Group 9 commented on how they were not sure how to integrate sensory cues into their assessment process. "*Humidity and all those things probably affects the environment...like the temperature and everything...[but] maybe we weren't expert enough...to understand that.*"

Developing Intuition

Developing an intuition for the training assessments using the reference images was the ultimate goal of the training. Participants had trouble remembering the reference images, and differed in their ability to integrate their individual assessments into a larger score.

Several groups vaguely recalled the training, but had trouble explicitly thinking of the reference images. For instance, [7A] and [7B] said that they had an intuitive "*kind of an idea...of the reference*" but "*[weren't] thinking about those images.*" Similarly, [10A] said that the training created intuition about the assessment task because it was "*much more like an experience rather than some memorization,*" however the pair recalled that "*it [took] more additional steps for me...to remember what the reference images were.*"

Participants differed in their comfort integrating the individual feature assessments into a larger score. For instance, [3A] felt the outdoor rating "*was fairly easy given that the images we saw on the Desktop was pretty similar to what we see [outside]...you don't have to collect all of the information from scratch...you just have to adjust it a little.*" In contrast, [8A] did not know how to integrate feature assessments. "*When I gave it the final score, I was like, well this side of the left bank is really rough...but this was really gentle.*" they explained. "*I've given it an average...[but that] might not be a good way*" to integrate the ratings together.

Training Outcomes

Going through the process of developing expertise improved participants' confidence and preparedness for the task, generated task enjoyment, and helped participants develop an appreciation in outdoor habitats. Many participants felt prepared for the assessment, or confident that they would get better over time. For instance, [4A] "*felt very prepared...pretty confident...how to rate,*" even commenting that although "*more...training helps...I'm not sure...if I would necessarily get better.*" [8A] also felt that their assessment skills improved as a result of training. "*I thought the training helped,*" then explained "*because if you would have taken us [outside] and tried to teach us, like standing there, I've been like, what are you talking about? But it was like, we had done a few repetitions of it, so we were more like, we knew a little bit more.*" Likewise, [5A] and [5B] "*definitely [felt] better prepared...than before...training,*" but remarked that "*if I were to...actually...collect data on the condition of the bank I would want to be better trained.*"

As well as generating confidence, the training created learning enjoyment for some participants. [8A] said that VR and multisensory realism was "*was fun. I really liked the whole environment,*" and Group 9 agreed that "*its nice to see [what made streams] good...why it's healthy.*" [2A] explained that they "*liked the co-learning experience because...because we both kind of grew in it together...it made it honestly more enjoyable...and made me want to advance.*" However, not all groups equally enjoyed the learning experience. When asked about working with a partner in future training, [8A] and [8B] offhandedly remarked that "*I don't know...[if I] would voluntarily just do this...you know?...is it my vocation?*"

For some participants, task enjoyment translated into a larger interest and appreciation for outdoor habitats. At the end of training, [2A] commented that the training "*was interesting...I'd never done anything like that before...and...it's stuff that we see in our everyday life...so it's kind of cool that we're trying to evaluate it.*" [3A] was also "*really excited to see how the knowledge I learned, I just learned, applied to the real world.*" They further commented that as a result of training, they would start noticing outdoor streams. Notably, Group 7, who completed the assessment task several hours after training, [9] said that streams were "*not something I...[pay attention to]...but after we did the [training] stuff...in the morning...I noticed [them] a lot more*" over the course of the day.

7.4.1.7 Summary of Qualitative Results

The qualitative thematic coding identified 4 themes expressed through 26 individual codes. The *Training Framework* theme found that participants found the system easy to use, but identified several needs; participants wanted higher quality resolution, and wanted reference images to better match the outdoor seasonality. Participants also identified interaction features that could improve training, such as having a way to point to features in VR. In complement to this, the *Immersion* theme found that study participants were engaged by the VR experience. In the multisensory condition, some participants found the additional sensory information engaging and helpful, whereas other did not.

The *Personal vs. Collaborative Experience* theme found that participants enjoyed and benefited from working with a partner. Collaborating with a partner

[9]The participant pair came late to the morning study session. They finished the outdoor data collection after their classes that afternoon.

helped participants better understand habitat features, and calibrate their decisions. This theme also uncovered that partner relationships differed; some wanted feedback during the entire training and assessment task, while others wanted to make evaluations on their own, and use their partners to check their answers.

The final theme, *Developing Expertise*, integrated the first three themes into participants' internal learning process. Many participants began training expecting it to be difficult, but soon learned to follow procedural tasks. As they assimilated environmental patterns, they began to construct a sense of intuition for assessment, and because more confident and interested in water monitoring. The training provided participants with an effective foundation, however the qualitative results found that participants required further training to fully develop a sense of intuition.

7.4.2 Quantitative Results

In this section, I report statistical differences between the three study conditions, StreamBED Standard VR, StreamBED Multisensory VR, and PowerPoint (PPT). Across these conditions, I consider differences in 1) training engagement, 2) training interaction, 3) outdoor engagement and interaction, and 4) outdoor assessment scores.

First, I present summary statistics for *training engagement* measures for Standard and Multisensory VR StreamBED training conditions, then summarize engagement measures for PPT participants. After summarizing individual metrics, I compare all three conditions on relevant measures. I likewise present parallel statistics and comparisons for *training interaction* measures. After describing training engagement and interaction, I summarize and compare participant *outdoor assessment engagement, interaction, and scores* across all three conditions.

To enhance readability, this section presents average means and significant differences between conditions. Descriptions of study measures and correlations are in Appendix A.

7.4.2.1 Summary of Engagement

This section first describes participant engagement measures in the Standard and Multisensory VR conditions, then describes engagement measures in the PPT condition. Overall, all three participant groups were immersed and engaged by the different training experiences.

Evaluated on a 7-point Likert scale, participants in both VR conditions were somewhat immersed by training (*Standard* (μ = 6.0, s = .666), *Multisensory* (μ = 4.9, s = 1.874)). Likewise, participants in the two conditions felt training was moderately realistic (*Standard*(μ = 4.9, s = .994), *Multisensory*(μ = 4.5, s = 1.5)), and that the training environment was relatively familiar (*Standard*(μ = 5.1, s = 1.370), *Multisensory*(μ = 5.1, s = .737)).

Both participant groups were highly engaged by VR training (*Standard*(μ = 5.7, s = .949), *Multisensory*(μ = 6.2, s = 1.229)), and reported being moderately engaged as the VR partner (*Standard*(μ = 4.9, s = .1.449), *Multisensory*(μ = 5.5, s = 1.443)).

PPT participants were likewise engaged by training. Measured on a 5-point Likert scale, PPT participants found the training very realistic ($\mu = 4.307$, s = .751) and familiar ($\mu = 3.923$, s = 1.256). Further, PPT participants were somewhat engaged by their training experience ($\mu = 3.385$, s = .1.044).

7.4.2.2 Differences in Engagement Measures

This section first considers differences between Standard and Multisensory VR conditions, then compares engagement measures between StreamBED and PPT conditions. Overall, participants in the StreamBED and PPT conditions were equally engaged by training, however PPT participants found training to be more realistic. Standard and Multisensory VR participants found the training to be equally immersive (x^2= 2.64, p = .104, df = 1), realistic (x^2=.034, p = .852, df = 1), and equally familiar (x^2= 0.067, p = .796, df = 1). Likewise, both VR groups were equally engaged by VR training (x^2= 2.639, p = .104, df = 1), and felt equally engaged in the role of the VR partner (x^2=.601, p = .438, df = 1).

Comparisons across the three conditions were measured on the converted 5-point Likert scale described in Figure 7.2. PowerPoint participants were as equally engaged as VR participants (x^2= 3.574, p = .167, df = 2), and found the training environment to be equally familiar (x^2=.760, p = .684, df = 2). Surprisingly, PPT participants found training more realistic than VR participants (x^2= 6.966, p = .03, df = 2). One explanation for this discrepancy may be that PPT participants could have misinterpreted "environment realism" to mean how realistic it was to learn in a PowerPoint setting, not how realistic the PPT training environment was relative to the real world.

7.4.2.3 Summary of Training Interaction

This section first reports on participant interaction measures in the Standard and Multisensory VR conditions, then describes participant interaction in the PPT condition. Participants across all three conditions felt they could interact well with training.

Training Clarity

Evaluated on a 7-point Likert scale, participants in both StreamBED conditions rated VR training as clear and appropriate. Both conditions felt that VR training was clear (*Standard*($\mu = 5.9$, s = 1.101), *Multisensory*($\mu = 5.9$, s = 0.994)), and gave them clear feedback (*Standard*($\mu = 5.3$, s = 0.823), *Multisensory*($\mu = 5.9$, s = 0.994)). Participants further considered this feedback appropriate for the VR training task (*Standard*($\mu = 5.4$, s = 0.699)), *Multisensory*($\mu = 6.3$, s = 0.823)).

Participant also rated the Desktop training as clear and appropriate. StreamBED participants thought the Desktop training was considerably clear (*Standard*($\mu = 5.2$, s = 1.033), *Multisensory*($\mu = 5.9$, s = .994)), and gave them clear feedback (*Standard* ($\mu = 5.3$, s = 1.059)), *Multisensory*($\mu = 5.5$, s = 1.176)). As with VR training, participants felt that Desktop training provided appropriate training feedback.

Interaction

Participants not only thought the VR training was clear, but also felt they could easily interact with training (*Standard*(μ= 4.5,s = 1.779), *Multisensory*(μ = 5.0, s = 1.633)), and could collaborate well with their partner (*Standard* (μ = 4.8, s = 1.229), *Multisensory*(μ = 4.4, s = 1.84)). Although they could easily interact with training, Standard condition participants felt somewhat overloaded with information during the VR training task (μ = 3.2, s = 1.476), notably more than Multisensory participants (μ = 2.0, s = 1.155). In parallel to this, VR partners felt they could contribute better to assessment in Multisensory VR (μ = 5.1, s = 1.287) than in the Standard condition (μ = 3.6, s = 1.646).

Although the Desktop training was clear, participants in both StreamBED conditions considered training relatively difficult to interact with (*Standard*(μ = 2.6, s = 1.506), *Multisensory*(μ = 2.7, s = 1.767)), and felt somewhat overloaded with information (*Standard*(μ = 3.4, s = 1.506), *Multisensory*(μ = 2.2, s = .476)). One explanation for this is that making qualitative feature assessments is a more difficult task than taking snapshots and tagging them. A contrast to the VR section findings, participants in both groups felt they could only moderately contribute to training as the Desktop partner (*Standard*(μ = 3.0, s = 1.700)), *Multisensory*(μ = 3.2, s = 1.874)).

The PowerPoint condition was a passive training, so participants were not asked about their ability to interact with the system. Instead, participants were asked how well they could contribute to the training experience during group discussion. On a 5-point Likert scale, participants felt they could contribute to the discussion relatively well (μ = 4.077, s = 1.115).

7.4.2.4 Differences in Training Interaction

There were few significant differences between StreamBED conditions during the VR training. Both groups rated the VR training and Desktop training as equally clear. Likewise, both StreamBED conditions were equally comfortable interacting with training and with their partner.

Notably, Multisensory participants felt marginally less overloaded by VR training than Standard participants (x^2= 3.258, p = .071, df = 1) and rated VR feedback as significantly more appropriate (x^2= 5.1831, p = .029, df = 1). Further, as the Multisensory VR partner, participants felt they could contribute significantly more to the training experience (x^2= 3.752, p = .052, df = 1).

As with VR training, Multisensory participants were marginally less overloaded by the Desktop experience than Standard participants (x^2= 3.166, p = .076, df = 1), however both groups rated Desktop training feedback equally (F(1,18) = 1.037, p = .322)), and both conditions similarly contributed as the Desktop partner (F(1,18) = .063, p = .805))

Multisensory participants being less overloaded during VR training may be explained by sensory information providing context and information coupling through ambient sensory cues [46, 74, 52]. However, this finding should be considered against the fact that Multisensory condition partners had a more active role during VR training; even if the ambient cues added value to participants' experience, Multisensory VR participants may have required more support from their partner. During Desktop training these sensory cues were not present, so participants may have not required additional partner support.

There were no significant differences between Standard and Multisensory participants' ability to interact during training (x^2= 2.143, p = .143, df = 1), however, participants in the VR condition felt they could contribute to training significantly better than PPT participants (x^2= 4.613, p = .032, df = 1). This is logical given that the VR training was meant to be a dynamic experience, whereas the PPT training was more passive. In contrast, participants' ability to contribute to training differed between the StreamBED and PPT conditions. There was a marginal significant difference between the three conditions (x^2= 5.138, p = .077, df = 2).

7.4.2.5 Outdoor Assessment

This section reports on the outdoor assessment experience all participants completed. I describe and compare engagement and interaction experiences across the three conditions, then compare participant scores to an expert habitat assessment.

Outdoor Engagement and Interaction

Overall, all participants were engaged by the outdoor assessment. Study results show that participants across all three conditions enjoyed the outdoor assessment (*Standard VR*(μ = 5.3, s = 1.947), *Multisensory VR*(μ = 6.2, s = .919), *PPT*(μ = 5.076, s = 1.038)). Although they enjoyed it, participants across conditions found outdoor assessment moderately challenging (*Standard VR*(μ = 4.4, s = 1.647), *Multisensory VR*(μ = 4.1, s = 0.994), *PPT*(μ = 4.231, s = 0.927)), and relied on their partner during the assessment (*Standard VR*(μ = 4.3, s = 1.337), *Multisensory VR*(μ = 4.4, s = 1.350), *PPT*(μ = 4.615, s = 1.044)).

When asked about the future, participants across the conditions expected to enjoy future data collection (*Standard VR*(μ = 4.7, s = 1.337), *Multisensory VR*(μ = 4.2, s = 1.476), *PPT*(μ = 4.077, s = 1.188)). However, VR participants were more excited for future training (*Standard VR*(μ = 5.0, s = .816), *Multisensory VR*(μ = 5.4, s = 1.075), than PPT participants(μ = 3.769, s = 1.481)).

Differences in Outdoor Engagement and Interaction

There were significant differences between VR and PPT participants on assessment enjoyment, and on excitement for future training. Multisensory VR participants enjoyed outdoor training significantly more than PPT participants (x^2= 5.765, p = 0.016, df = 1), but enjoyment was not significantly different between VR conditions. Also notable is that the three study conditions expected to enjoy future data collection equally, however, participants in both VR conditions expected to enjoy future training more than PPT participants (*Standard VR* (x^2= 6.59, p = 0.010, df = 1), *Multisensory VR* (x^2= 7.481, p = .006, df = 1)). This finding suggests that VR participants were more interested in understanding habitats better through training.

Assessment Scores

Participant outdoor scores for Epifaunal Substrate and Bank Stability were compared against a “gold standard” that I made that was verified by an expert. Difference between scores were calculated as the absolute value of a participant’s

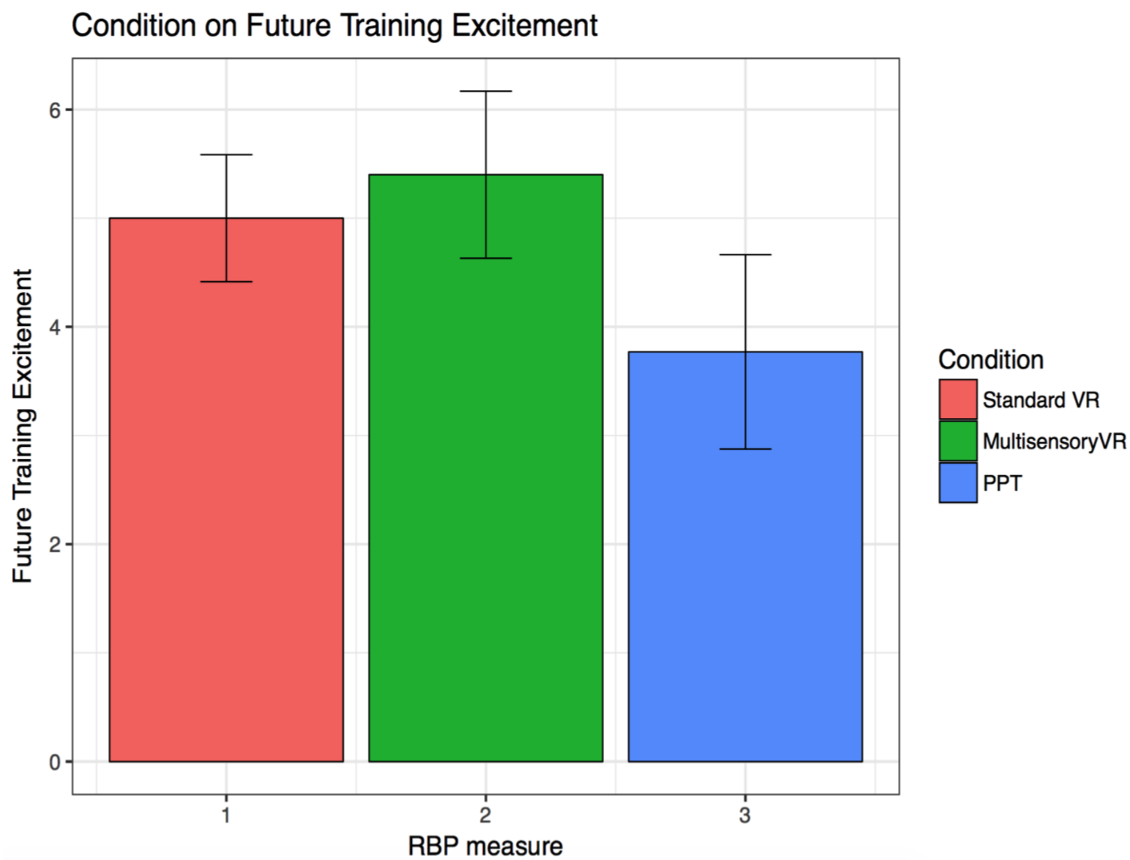

Figure 7.20: Bar chart showing the differences in Future Training excitement across the three conditions. Participants in the Standard VR (red) and Multisensory VR (green) conditions expected to enjoy future training significantly more than PPT participants (blue).

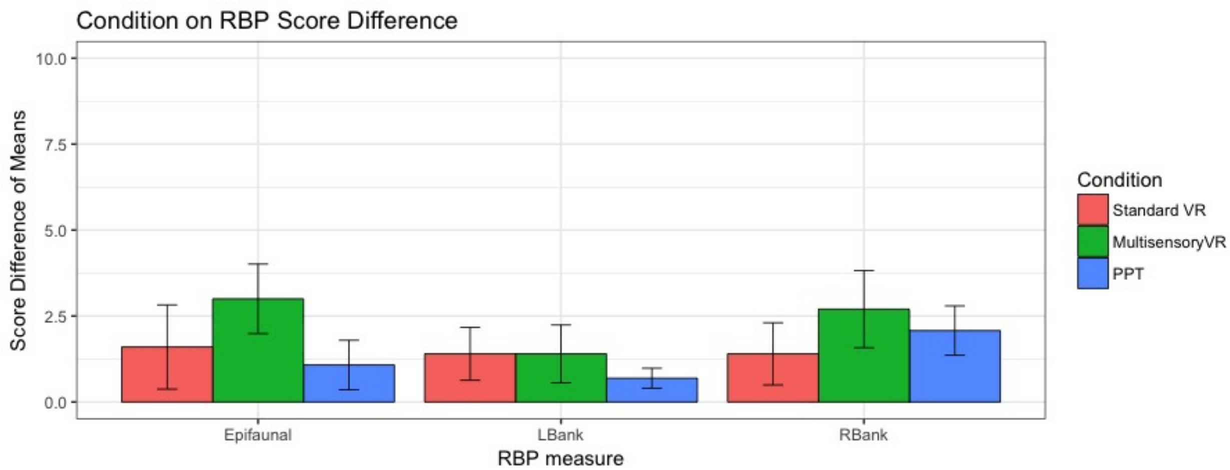

Figure 7.21: Bar chart showing the absolute difference between participant and the gold standard scores (a score of "0" indicates that participants had the same score as the gold standard). For the Epifaunal Substrate measure, participants in the Standard VR (red) and PPT (blue) conditions scored significantly closer to the gold standard than Multisensory VR participants (green). For the Left Bank Stability measure, participants in the Standard VR and PPT conditions likewise scored significantly closer to the gold standard than Multisensory VR participants. For the Right Bank Stability measure, participants in the Standard VR condition scored significantly closer to the gold standard than Multisensory VR participants. The study data was not normally distributed, so the Kruskal-Wallis, a non-parametric test, was used to compare conditions.

score subtracted from the expert score. As described in chapter 2, RBP scores for the two metrics are rated differently; Epifaunal substrate are evaluated on a 20-point scale, while scores for Bank Stability are evaluated on a 10-point scale. For this reason, larger differences on Epifaunal Substrate should be considered in proportion to smaller differences on Bank Stability.

When rating Epifaunal Substrate, PPT participants scored closest to the expert score, (μ = 1.077, s = 1.188), and Standard VR participant scored almost as close (μ = 1.6, s = 1.713). Multisensory VR participants scored farther from the expert score, on average (μ = 3, s = 1.414).

When evaluating Left Bank Stability, PPT participants likewise scored closest to the expert score, (μ =.692, s = .480). On this metric, Standard VR participants (μ = 1.4, s = 1.075) scored similarly to Multisensory VR participants (μ = 1.4, s = 1.174), but with a smaller variance.

In contrast, when evaluating Right Bank Stability, Standard VR participants scored closest to the expert (μ = 1.4, s = 1.265), while Multisensory VR (μ = 2.7, s = 1.567) and PPT participants (μ = 2.077, s = 1.188) scored further from the expert.

Comparing Scores Across Conditions

Comparing participant Epifaunal substrate and Bank Stability scores across the three conditions, I found significant differences between Standard VR, Multisensory VR and PPT participants on the different scores. On Epifaunal Substrate, there was no significant difference between Standard VR and PPT participant scores

(x^2=.609, p = .4351, df = 1), however, participants in the Standard (x^2= 5.153, p = .023, df = 1) and PPT condition (x^2= 8.105, p = .004, df = 1) scored significantly closer to the expert score than Multisensory VR participants. Likewise, on Left Bank Stability, participants in the Standard (x^2= 4.008, p = .046, df = 1) and PPT condition (x^2= 3.910, p = .046, df = 1) scored significantly closer to the expert score than Multisensory VR participants. [10]

In contrast, when comparing conditions on Right Bank Stability, there was no difference between Standard VR and PPT conditions (x^2= 1.972, p = .160, df = 1), and no difference between Multisensory VR and PPT conditions (x^2= 1.177, p = .278, df = 1). However, Standard VR participants scored significantly closer to the expert score than Multisensory VR participants (F(1,18) = 4.856, p = .04)).

Score comparisons show that the Standard VR and PPT conditions scored closest to expert scores, suggesting they were the most effective training conditions. However, an important consideration is the normality and spread of participant scores across conditions. As shown in Figure 7.22, participant scores were not normally distributed in all three measures, making it difficult to draw conclusions about the effect of conditions against the effect of individual differences between conditions.

7.4.2.6 Summary of Quantitative Results

The quantitative results identified differences in participant engagement, interaction, and scores during training and outdoor data collection. Overall, all three

[10] Participants in the Standard and Multisensory VR condition had the same mean, however, the data was not normally distributed. Instead of comparing means, a more complex non-parametric test was used to determine significance.

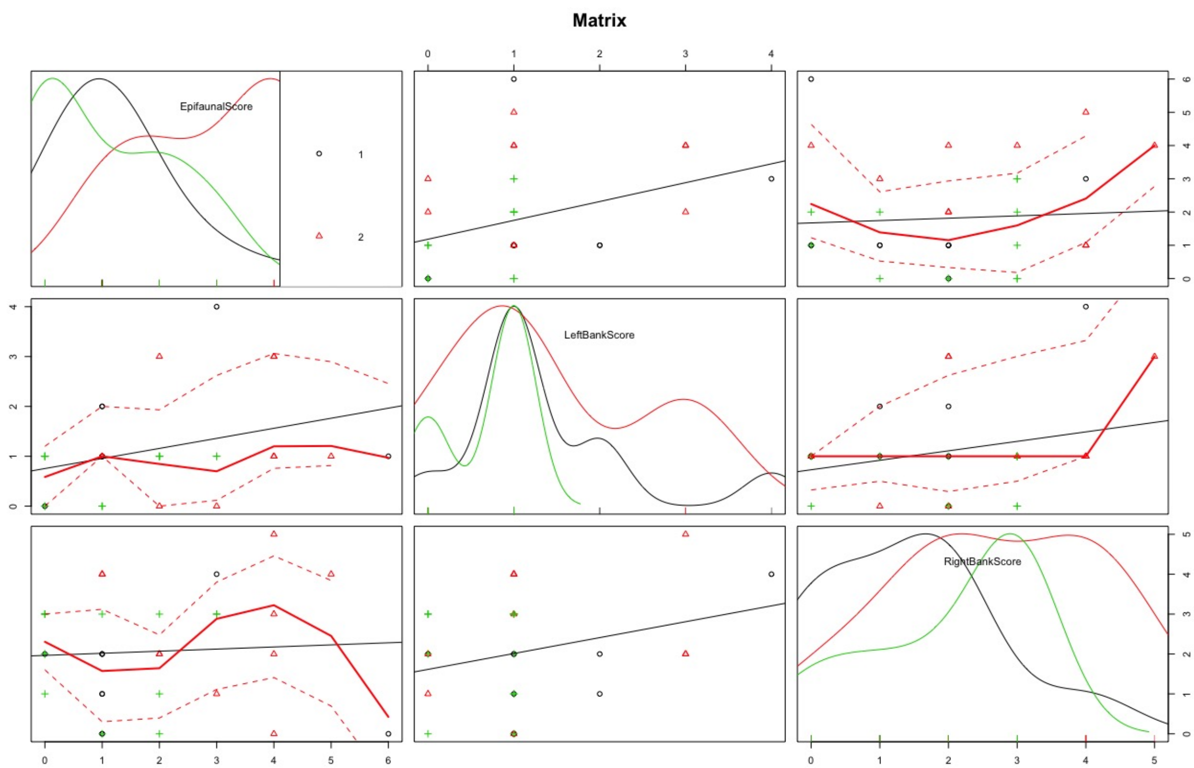

Figure 7.22: Scatterplot matrix showing differences in participant conditions for Epifaunal Substrate, Left Bank Stability and Right Bank Stability. The Standard VR condition is in black, the Multisensory VR condition is shown in red, and the PPT condition is in green.

participant groups were engaged by training, however, differences in outdoor data collection scores suggest that Standard VR was the most effective condition because participants on average scored closest to expert assessments, and were more excited by future training.

When comparing outdoor assessment scores to the gold standard score, Standard VR participants made assessments closest to expert scores, while Multisensory VR participants made assessments that were furthest from the gold standard. When evaluating the 20-point Epifaunal Substrate scale, Standard VR and PPT participants were on average 1.7 points (8.5%) closer to the gold standard than Multisensory VR participants. Likewise, when evaluating the 10-point Bank Stability scale, Standard VR and PPT participants were on average .354 points (3.5%) closer to the Left Bank expert metric than Multisensory participants. When measuring Right Bank Stability, Standard VR participants were 1.3 points (13%) closer to the gold standard score than Multisensory VR participants. Standard VR and PPT participants performed significantly better than Multisensory participants, however rating differences between conditions were remarkably small.

In complement to this, the study found that participants in both VR conditions expected to enjoy future training more than PPT participants. Although both groups were interested in the data collection task, this finding suggests that VR participants were interested in learning more about stream habitats and getting better at qualitative assessment.

7.4.3 Summary

The goal of this study was to test the effectiveness of the StreamBED VR system on ability to learn to make qualitative habitat assessments. In this chapter, I described the iterative process of building the StreamBED VR training system. Then, I tested two versions of the prototype (Standard VR and Multisensory VR) against PowerPoint training, a baseline similar to standard citizen science RBP protocol training, pairing bottom-up thematic analyses with quantitative metrics of engagement, interaction, and difference in scores.

My qualitative analyses uncovered four important themes; *Training Framework*, *Immersion*, *Collaborative Decision Experience*, and *Process Developing Expertise*. Overall, I found that the training framework effectively scaffolded the learning process, and that participants were immersed and engaged by the system. Further, I found that participants benefited from working with a partner during training, and that their expertise developed from the beginning to the end of training. Further, participants began to develop intuition at the end of training, however, they needed more time to remember reference habitats and to build intuition.

My quantitative findings echoed many of my qualitative findings; participants across both VR conditions rated the system as clear and easy to interact with. Comparing training effectiveness across all conditions, the quantitative findings revealed that Standard VR participants made assessments significantly closer to gold standard scores than Multisensory VR participants, but also that PowerPoint participants scored equally well as Standard VR participants. Although these groups scored equally well relative to the expert standard, the data suggests that VR participants were more interested in future training more than PPT participants, indicating they were motivated to get better at the qualitative assessment task.

Chapter 8: Discussion

8.1 Summary of Contributions

My first study found that experts learned to make assessments by experiencing a range of habitats through an implicit process; they developed intuition by interpreting a range of visual and sensory cues not part of explicit protocol guidelines using peer discussion and negotiation of meaning. Likewise, the StreamBED pilot study found that VR could replicate learning habitats, but that citizen science learners had trouble making judgments of VR training environments that did not match real world visual stimuli and interaction, and that they needed the expert learning process to be scaffolded into clear and actionable tasks. In tandem with this, my multisensory realism study found that ambient sensory cues improved observation skills in VR, but their value was qualified by practical setup considerations.

The goal of the final StreamBED VR design was to integrate findings from the first three studies into a realistic habitat training prototype; this training scaffolded the qualitative assessment learning experience with collaborative learning. My iterative design process found that trade-offs existed in VR between supporting realism and interaction, and that my design process was subject to changing VR system affordances.

Based on this training, my final study exposed the complexity of externally scaffolding the expert internal learning process; I found that VR was useful for building intuition for physical environments, but that learners need time to internalize complex relationships between habitat features. I also discovered that asymmetrical collaborative learning helped scaffold training well; peer learning helped participants become comfortable with training material and gave them confidence.

Table 8.1 overviews these key research contributions. My research shows that while VR has the potential to help train judgment tasks, training design is multifaceted and complex, full of theoretical learning considerations and practical challenges. Further, VR realism can be a powerful tool for training, however, it is only effective when the training experience fully parallels the assessment task.

8.2 Individual Studies

The following section summaries each study, and considers individual findings in context of related literature.

8.2.1 Study 1: Differences in On-site Experiences

The goal of the first study was to assess the viability of training citizen science volunteers to make qualitative assessments of streams and watersheds using the

Table 8.1: Key Research Contributions

Study	Domain	Finding	Contribution
#2,4	*Research through Design*	VR realism and interaction have trade-offs that rapidly change with system innovation.	In HCI, the RTD process must adapt to advances in technology state-of-the-art. The process should consider how designers should adapt requirements to changing system affordances.
#2,4	*VR Realism*	Learners have trouble making judgments of VR training environments that do not match real world visual stimuli and interaction.	VR training environments must replicate training stimuli realistically. Differences in visual stimuli or interaction can negatively affect training.
#3,4	*VR Realism*	Multisensory ambient cues can improve observation skills in VR. However, their value may change based on training context.	Ambient sensory information can add value to learning experiences, allowing learners to form more complex environment relationships. The value of ambient cues should be considered within the scope and complexity of training design.
#2,4	*VR Training*	VR can help build intuition for physical environments, but learners need time to internalize complex relationship dynamics.	VR may be an effective tool for training assessment of physical environments, particularly when actual environments are impractical or dangerous to experience in person. Intuition judgment training must be done over time.
#1,4	*Citizen Science*	Experts develop intuition for stream monitoring using environmental factors that are not part of official protocol metrics. Scaffolding this intuitive process into an explicit framework is challenging.	When developing citizen science training, HCI researchers must consider that experts may augment judgment or reasoning with implicit sources of information. The expert process may not directly translate to training, and must be made explicit to novice learners.
#1,4	*Peer Learning*	Asymmetrical VR training can create effective peer learning. Such training encourages negotiation of meaning, and helps learners feel comfortable with new material.	Such asymmetrical training can help designers and educators create economical shared experiences in VR. Training design should support learner discussion, and allow learners to experience both partner roles.

EPA's Rapid Bioassessment Protocol (RBP) [6]. Through my research, I examined the challenges of interpreting the RBP, and discerning between professionals and volunteer water monitor experiences. My research found that the biggest difference between professional and volunteer monitors were volunteers' lack of experience with on-site streams; professionals experience many different streams throughout their career, whereas volunteers only see a handful of streams, likely in their immediate neighborhood or community. My findings suggest that professionals use their experience to make intuitive habitat judgments that extend beyond RBP measures: they interpret stream characteristics using ambient sensory information, and are able to describe past, present, and future narratives of different streams.

That professional monitors make intuitive judgments can be at least partially explained by the challenges of making qualitative assessments using the RBP: the protocol asks professionals to subjectively interpret differences in quality based on their experiences and to account for misleading measures. I found that experts often formed these intuitive judgments through narrative; rather than making isolated assessments, many professionals interpret the state of a stream as part of a changing narrative. This suggests that professionals develop higher-order thinking skills, inductively envisioning the stream's history and future based on their knowledge of how stream features affect one another. This finding strongly echoes discussed literature on scientific learning [8, 118]. This is also in line with Bloom's education taxonomy [73], which suggests that beyond analyzing and evaluating information, students should be able to reassemble information into new ideas. Further, that professionals scaffolded their intuitive knowledge through environmental sensory information confirms findings from Psotka [104] and Dihn [32], who suggest that multisensory cues can reduce conceptual load, create salient memories and emotional experiences, and increase memory and sense of presence for environmental information.

This study also found that experts learn to interpret intuitive judgments through peer discussion and feedback. This is in line with Crossan's work on creating shared interpretations of meaning [20]. This also follows Flyvbjerg [39], who explains that self-assessment may be improved by benchmarking choices and performance against others, and with Eberbach and Crowston [34], who examine the effect of discussion on reviewing and updating internal information scaffolding. These findings suggests that effective training should include a feedback loop that allows learners to discuss and receive feedback from peers or teachers.

8.2.2 Study 2: Learning from the VR Pilot

The goal of the second study was to consider the viability of the StreamBED VR initial training design, and to understand novice learner needs. Together, the quantitative and qualitative findings suggest that the virtual training tasks produced unique learning opportunities that engaged and motivated study participants; the quantitative data revealed a positive shift in immersion and motivation, while qualitative trends suggests that participants went of their way to interact with the training by exploring the environment and telling stories. Participants increased sense of immersion and motivation as a result of VR presence is reflected in literature. For instance, Dede [28] and Bryanton [14] suggests that immersion resulting from VR presence is intrinsically helpful to developing motivation, allowing learners to practice mastery of complex knowledge and skills.

Study findings also revealed system design challenges; several participants did not fully understand or pay attention to the protocol. Game literature suggests that to be effective, users must have clear goals [87, 35]. McMahan [87] further clarifies that players should have meaningful interactions with non-trivial impact on the environment. The study further suggested that participants had trouble interpreting and making assessments in the rendered VR environment. This contradicts McMahan [87], who explains that in games, total photo and audio realism is not necessary to create immersion. Unlike McMahan, however, the goal of the pilot study was for participants to interpret real world habitats based on VR training. In this case, the disconnect came from comparing the detail of a rendered environment to the real world.

8.2.3 Study 3: Value of Multisensory Information

The third study considered the role of ambient multisensory cues on the number of observations and inferences that participants made, and on their immersion in the VR stream habitats. My findings reveal that participants across conditions make observations and inferences using multiple senses; when given more sensory information, participants were likely to use them to better understand habitats.

My grounded theory findings suggest that both groups synthesized sensory information into their observations. For instance, most participants observed a scream from a mountain lion in the second video, and participants across conditions included sensory information in their storytelling. In tandem with this, my quantitative findings reveal that participants in the Multisensory VR condition made significantly more visual and auditory observations than Standard VR participants. This suggests that ambient multisensory information may have augmented participants capacity to understand and describe their environment.

These finding are supported by cognitive neuroscience research suggesting that congruent multisensory stimuli coactivates parallel sensory channel information processing [46, 74]. This is likewise supported by design research suggesting that synesthetic design leads to cross-sensory information coupling [52]. The study found that multisensory information added value to immersive learning environments, allowing participants to make more detailed observations and remain engaged. However, this value must be qualified by context; Winn [125] suggests that the extra cost of high immersion only pays off when the content to learn is complex and three-dimensional.

8.2.4 Study 4: Evaluating StreamBED 2.0

The final study integrated findings from the first three studies into the StreamBED VR 2.0 training. The research tested two versions of the prototype (Standard VR and Multisensory VR) against PowerPoint training, a baseline similar to standard citizen science protocol training [126]. The study paired bottom-up thematic analyses with quantitative metrics.

The qualitative and quantitative findings revealed that participants across all three conditions found training relatively clear and engaging. In VR, participants that were worried about training difficulty were surprised to find how easy it was, and became comfortable with the scaffolded process of making assessments. The outdoor quantitative measures found that the three groups equally enjoyed data collection, and equally expected to enjoy future data collection. However, both VR groups expected to enjoy future training significantly more than the PPT group (5.0 vs. 3.767 on a 7-point Likert scale), implying that VR participants were more interested in the learning task.

Comparing scores across the three conditions yielded surprising results. Both Standard VR and PowerPoint participants made assessments that were significantly closer to expert scores than Multisensory VR participants; on the 20-point Epifaunal substrate scale, there was a 1.7 point difference between the Multisensory VR condition and Standard VR and PPT groups, and on the 10-point Bank Stability scale, there was a .354 point (left bank) and a 1.3 point (right bank) difference between these groups.

A remarkable finding of this study is that participants in the baseline PowerPoint condition performed equally well to Standard VR participants, even though the PPT training was less interactive and took half as much time (60 minutes vs. 135 minutes). One explanation for this may be that StreamBED participants learned to use the VR technology during training, spending on average 20 minutes interacting with the VR tutorial. This additional learning task may have overwhelmed them with information. Participants described the training as intuitive, however, quantitative results show that participants in both VR conditions were moderately overwhelmed (2.6 on a 7-point Likert scale) by the training interaction.

Another explanation for this finding is that participants found value in the passive PowerPoint experience. HCI literature suggests that active interaction [24] and personalization [25] are beneficial for learning, however, the Desktop interaction may have hampered participants' ability to identify patterns between reference images; PowerPoint participants saw the full set of reference images before making assessments, whereas StreamBED participants immediately began the feature assessment task using the reference images. Without first recognizing similarities and differences between the range of features [20], StreamBED participants may have had trouble synthesizing the reference images into patterns [3, 122].

It is also striking that the Multisensory condition performed worse than both the Standard VR and PPT conditions, particularly because Multisensory participants were marginally less overloaded by training than Standard VR participants (2.0 vs. 3.2 on a 7-point scale). One explanation may be that the multisensory media simultaneously supported and impaired participants' experience. Described in study 3, multisensory stimuli activates parallel information processing [46], which may have allowed Multisensory participants to more easily orient themselves in the 360° habitats. However, because the ambient stimuli did not directly correspond to the RBP protocol measures, it may have distracted participants from the assessment tasks. As an example, Group 9 tried to connect the ambient sensory information to assessment, remarking that "*humidity and...temperature...probably affect the environment...[but] maybe we weren't expert enough....to understand [how].*" The mulsemedia meant to enhance the realism of the training environment, however participants incorrectly interpreted the stimuli as cues with which to evaluate the protocols.

Finally, although Standard VR and PPT participants outperformed the Multisensory condition, the difference in ratings between the three conditions was appreciably small (on average 1.3 points on the 20-point Epifaunal scale, and .827 points on the 10-point Bank Stability scale). Study 1 found that expert monitors do not always agree about assessment scores, and Roth [113] confirms that small score variations are accepted in the qualitative water monitoring community. The quantitative data is too small to account for individual differences, so it is difficult to ascertain their effect on condition. Participant scores were not normally distributed, and the scatterplot for the Right Bank Stability measure (with the largest difference between the gold standard and participant scores) shows a large spread in the data, suggesting the effect of other factors. This is anecdotally reflected in the data; a Multisensory VR participant casually disclosed that their mind was not "neurotypical," and that they may have responded differently to training stimuli than other participants.

Participants felt comfortable with the assessment process, and the thematic codes revealed that they began to develop some intuition for assessment. However, participants required more time to remember reference habitats, and to incorporate them into their judgment process. This finding reflects literature on qualitative learning. Qualitative judgment skills develop over time, and are difficult to examine and compare [20]. This finding also reflects psychology literature suggesting that sleep helps people process learned information [97, 116], and education literature showing that learners need time to process information into patterns [111]. The complexity of this research makes it difficult to gauge the relative value of VR training and PowerPoint training. Future research should expand on this work though iterative training and testing tasks.

This study suggests that both VR and PowerPoint are valuable tools for training qualitative judgment tasks. I found that VR has the power to engage, immerse, and motivate learners. However, these tools also have the ability to overload and distract learners, and may not be effective for all audiences. Surprisingly, I found that PowerPoint can also train qualitative judgment tasks effectively, although it may not equally engage learners.

The goal of this research was to consider how technology can be used to train scientific qualitative judgment tasks. This chapter summarized my research contributions and contextualized four studies in the research questions and literature. The final chapter summarizes my research, and presents limitations and future work.

Chapter 9: Summary, Limitations, and Future Work

9.1 Summary of Research

In this dissertation, my research considers how technology can be used to train scientific qualitative judgment tasks. I ask this question in the domain of citizen science water monitoring, considering the expert assessment process, the value of realism, and the challenge of scaffolding expert learning tasks for citizen scientists.

First, I asked *how expert monitors make qualitative assessments*, to understand how their process may be transferred to citizen science training. Study 1 found that professionals learn to make assessments on-site, through peer discussion. Further, experts develop intuitive judgments skills by supplementing misleading and subjective protocol metrics with ambient environmental information.

Then, I asked whether *multisensory cues help learners make habitat observations*, considering whether ambient sensory information benefits learners ability to understand and interpret their environment. Study 3 found that participants in the Multisensory VR condition made significantly more visual and auditory observations than Standard VR participants, and that ambient sensory information augmented participants' ability to understand their environment.

Assimilating my individual research components, my final research question asked how *training should scaffold learning to support the expert process.* The goal of this question was to understand how training design could transfer the observation and assessment skills of expert monitors to citizen scientists. My VR pilot tested an initial design of StreamBED training, finding that participants required explicitly defined tasks and goals, needed support interpreting subjective protocols, and wanted training realism comparable to outdoor assessment habitats.

I developed the final training, StreamBED 2.0, in response to findings about the experts' learning process, pilot study needs, and the effectiveness of multisensory realism. The final study found that the scaffold I created engaged and motivated learners across all three conditions, and helped them make qualitative assessments that were close to the expert gold standard. In the StreamBED VR conditions, my research found that collaborative learning positively contributed to participants' experience, but that the value of multisensory cues was unclear in context of broader training. Overall, the research found that the scaffold I developed was an effective way of training habitat assessment, both in VR and in PowerPoint.

9.2 Limitations

This dissertation was an initial exploration of how virtual reality may be used to train qualitative assessment in stream monitoring. The research was limited by technology, time, and application constraints described below.

This research was heavily limited by technology affordances and constraints. A large factor that affected my design process was frequently changing VR affordances, which made it unclear how to best to design for realism and interaction trade-offs, and for collaborative learning. A second technology limitation was VR footage resolution. I captured 360° videos in 4K definition, however, this translated to lower quality resolution in VR, making it difficult for some users to see stream details like underwater vegetation. Future research will be subject to different affordances and limitations as VR technology changes.

Training season and time were also limitation of this research. Experts primarily conduct habitat assessments in the summer, however, due to time constraints, the final study was conducted in the winter, thus making it challenging to assess seasonal habitat features like bank vegetation. Further, the reference images illustrated summer habitat features because it was not possible to find enough reference images taken in the winter. The final study was also limited by study training time. Participants completed training during one session, however intuition takes time to build. Future work should conduct training during the same seasons that experts make assessments, and should present training as a multistage learning experience.

Finally, the research was limited by domain and sample size. This research focused entirely on water monitoring, a familiar domain to participants who spend time outdoors. It is thus unclear how the research would generalize to unfamiliar citizen science domains. While the research was complex, individual components of this work may transfer across domains. For instance, peer learning is being explored across learning science domains (e.g. [76]), and Multisensory VR is being explored in HCI visualization [95]. Likewise, the research was limited by sample size. Due to the complexity of each study, training was conducted with a limited sample size of approximately n=10 per condition, insufficient for making nuanced statistical comparisons. Future work should expand on this research to consider the effect of training in larger sample sizes across familiar and unfamiliar domains.

9.3 Future Work

This section considers how my findings can be applied to expert and citizen science training. This section also considers how future work on VR realism and collaborative learning can build on this dissertation.

9.3.1 Leveraging VR to Support Qualitative Assessment Training

The goal of my work was to understand whether virtual reality could be used to train qualitative judgment tasks. My work focused on citizen scientists, however it can be applied to expert learning experiences, and to other domains that require environment judgments. For instance, VR habitat assessment can be used to make assessments about mountain top coal mining [102, 101], a domain in which in-person training can be dangerous. Further, this research can be used in climate change education; VR habitat training can allow learners to experience, rather than read about, the gradual effects of climate change [118].

This research may also be applied to other citizen science domains. Although my work focused on water monitoring, my findings may inform other citizen science projects like eBird, IceWatch, and the Clean Air Coalition, that report on changes in physical, ecological, and societal habitat variables [85]. A goal of these projects is to help the public understand and appreciate complex ecosystems, and to engage them in identifying problems and solutions. Training citizen scientists to interpret qualitative measures related to these projects could help support this goal. For instance, birdwatchers could learn to assess ecosystem habitats based on the presence of different birds. Likewise, ice and pollution monitors could learn to identify relationships between observable and climate factors.

9.3.2 Visual and Multisensory Realism

This research found that learners required visual realism during assessment training, however it unclear how much realism is necessary for effective learning. Future HCI research should consider how much visual realism is necessary for learners: *How much realism is necessary for learners to scaffold experiences into an effective cognitive map?* Likewise, is *training photo-realism always necessary?*

Research should also consider the role of multisensory realism. My findings on realism were mixed; my multisensory observation study found that mulsemedia improved participant observation, but my final StreamBED VR training study found Multisensory VR participants scored farthest from the gold standard. Since my research was an initial exploration of multisensory realism in training, several questions remain. First, multisensory research should consider the roles of sensory cues in training: *to be effective, must sensory information convey information that is directly relevant to training? Does ambient sensory information add value by* *increasing experience immersion?* Further, study training should consider when sensory realism should be introduced during training: *should it be part of the entire training experiences, or should it be limited to a particular part of training?*

9.3.3 Collaborative Learning in VR

This work can also be applied to collaborative VR training, a burgeoning HCI domain in the process of identifying interaction challenges [49, 51, 59]. My dissertation found that asymmetrical training was an effective collaborative learning solution. Future work can expand on this research by considering different elements of collaborative interaction. For instance, *what does turn-taking look like in asymmetrical VR training? Should one partner be in control of the experience, or should control be shared?* The research also found that study participants had trouble communicating habitat features they were looking at. Further research should ask, *how might systems design for joint attention to inform partners about what their peers are seeing or doing?*

Finally, future research should consider when different collaborative dynamics are appropriate and practical. The final study focused on peer learning, however other dynamics (e.g. expert/novice or group dynamics) may be more appropriate for different situations. Future research should ask, *when learners would benefit from expert guidance, peer dynamics, or group learning?* Further, research should consider if *live collaborators are necessary for learning? Can training employ system non-player characters (NPC's)?*

9.3.4 Summary

The goal of my dissertation was to consider how technology could train scientific qualitative judgment tasks to citizen scientists. To this end, my research presented four studies considering how experts learn to make assessments, how multisensory realism contributes to observation skills, and how training can scaffold the qualitative learning process. My study results found that the scaffold I created effectively trained qualitative assessment skills, but that this scaffold could be taught as effectively in PowerPoint as in VR. However, my research found that VR has an inherently different value; it can engage and motivate learners more than traditional PPT methods. Given the importance of water in society and climate change, and the great potential of citizen science research, this dissertation has the potential to create high impact in education technology.

Appendix A: Appendices

A.1 Expert Interview Questions

1. Do you use a variant of EPA's RBP protocol in your work? How? What is it used for?

2. How long have you doing assessment? How often do you go out? Where?

3. How did you first begin doing qualitative RBP assessments?

4. How did you learn to make do qualitative RBP assessments?
   a. Can you tell me about the experience?
   b. What training did you receive? How long was it? Was it informative? Did it teach you all the skills you need for the job?

5. What scientific background did you have before learning to do this work? Do you think having that background was important to making qualitative assessments of the RBP protocol?
   a. What kind of background/expertise do you think is important to do good quality assessment?

6. Process: How do you go about doing assessment?
   a. What multisensory cues do you use? Do you use the protocol during assessment?

A.2 Pilot Study Questions

StreamBED VR Pilot

(Pre-Training)

Participant Number

____

Are you...

- ◯ Male (1)
- ◯ Female (2)

Which category best describes your ethnicity?

- ◯ African American (4)
- ◯ Asian American (5)
- ◯ Hispanic (6)
- ◯ Native American (7)
- ◯ White/Caucasian (8)
- ◯ Other (9)

What's the highest level of education you've completed?

- ◯ High School (1)
- ◯ Bachelors Degree (2)
- ◯ Masters degree (3)
- ◯ PhD (4)

Page 1 of 16

How many citizen science projects are you familiar with?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

How many citizen science projects have you been involved in?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

Did you participate in data collection on a citizen science project?

- ◯ Yes (1)
- ◯ No (2)

Display This Question:

Did you participate in data collection on a citizen science project? = Yes

If so, for the project you were most involved in:

Display This Question:

If so, for the project you were most involved in: Is Displayed

How many times did you participate in data collection?

- 1 (1)
- 1-2 (2)
- 3-4 (4)
- 5+ (3)

Display This Question:

If so, for the project you were most involved in: Is Displayed

What types of training did you receive?

- Pamphlets (1)
- Online resources (2)
- Lectures (3)
- Game (4)
- On-site training (5)

Display This Question:

If If so, for the project you were most involved in: Is Displayed

How many hours of training did you receive?

- 0 (1)
- 1-2 (2)
- 3-4 (3)
- 5+ (4)

Display This Question:

If If so, for the project you were most involved in: Is Displayed

How comfortable do you feel collecting data?

Not Comfortable at all	Somewhat Comfortable	Very Comfortable

0	1	2	4	5	6	7

How familiar are you with water quality monitoring?

Not Familiar at All	Somewhat Familiar	Very Familiar

1	2	3	4	5	6	7

.

Have you ever participated in water quality monitoring data collection?

- ◯ Yes (1)
- ◯ No (2)

Display This Question:

If Have you ever participated in water quality monitoring data collection? = Yes

S2 If so,

Display This Question:

If so, Is Displayed

How many projects did you participate in?

- ◯ 1 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (4)
- ◯ 5+ (3)

Display This Question:
If so, Is Displayed

For the project you were most involved in:

Display This Question:
If so, Is Displayed

How many hours of training did you receive?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

Display This Question:
If so, for the project you were most involved in: Is Displayed

How comfortable do you feel collecting data?

	Not Comfortable at all		Somewhat Comfortable		Very Comfortable	
	0	1	2	4	5	6	7
Click to write Choice 1 ()							

Page 5 of 16

Display This Question:

If so, Is Displayed

How many times did you participate?

- ○ 1 (1)
- ○ 2-3 (4)
- ○ 4-5 (5)
- ○ 6+ (6)

[answer the following questions to the best of your abilities]

	Never		Sometimes		Always		
	0	1	2	4	5	6	7
I can understand all of the key concepts covered in my course. ()							
Soon after the end of a lesson, I am able to remember all of the key concepts ()							
Soon after the end of a lesson, I am always able to distinguish the most important concepts from concepts of less importance ()							
It is always easy for me to understand new information, even on a topic that does not interest me very much ()							
It is always easy for me to connect new information about a topic that interests me with other pieces of information ()							
When I find something new about a topic that I am studying, I am always able to connect it with other things that I know about the topic ()							
During a course, if we are given a new task to complete, I can always complete it by applying the knowledge that I obtained from lessons ()							

How many hours do you spend playing computer-based video games per week?

- ◯ 0 (1)
- ◯ 1-2 (5)
- ◯ 3-4 (6)
- ◯ 5+ (2)

How much experience do you have with an Xbox controller?

	Have never used it		I've used it a few times		I'm extremely comfortable with it	
0	1	2	4	5	6	7

Have you ever experienced virtual reality?

- ◯ Yes (1)
- ◯ No (2)

Display This Question:

If Have you ever experienced virtual reality? = Yes

How many times?

- ◯ 1 (1)
- ◯ 2-3 (2)
- ◯ 4-5 (3)
- ◯ 6+ (4)

Display This Question:

If Have you ever experienced virtual reality? = Yes

How comfortable did you feel interacting in a virtual reality environment?

	Not comfortable at all			Somewhat comfortable			Very Comfortable	
	0	1	2	3	4	5	6	7
Click to write Choice 1 ()								

[answer the following questions to the best of your abilities]

	Not at all		Somewhat		Very Much	
	0	1	3	4	6	7
How much do you expect to enjoy the training? ()						
How much do you expect to enjoy the data collection? ()						
How excited are you to interact with the training game? ()						
How excited are you to interact with the Virtual Reality headset? ()						

Additional Comments:

(Post-Training)

	Disagree / Neither Agree nor Disagree / Agree 0 1 2 4 5 6 7
I lose track of time ()	
Things seem to happen automatically ()	
I feel different ()	
I feel scared ()	
The game feels real ()	
If someone talks to me, I don't hear them ()	
I get wound up ()	
Time seems to kind of stand still or stop ()	
I feel spaced out ()	
I don't answer when someone talks to me ()	
I can't tell that I'm getting tired ()	
Playing seems automatic ()	
My thoughts go fast ()	
I lose track of where I am ()	
I play without thinking how I play ()	
Playing makes me feel calm ()	
I play longer than I meant to ()	
I really get into the game ()	
I feel like I just can't stop playing ()	

Page 9 of 16

[answer the following questions to the best of your abilities]

	Not At All			Partly			Very Much
	0	1	2	4	5	6	7
I feel just the right amount of challenge ()							
My thoughts/activities run fluidly and smoothly ()							
I don't notice time passing ()							
I have no difficulty concentrating ()							
My mind completely clear ()							
I am totally absorbed in what I am doing ()							
The right thoughts/movements occur of their own accord ()							
I know what I have to do each step of the way ()							
I feel that I have everything under my control ()							
I am completely lost in thought ()							

	Easy					Difficult	
	0	1	2	4	5	6	7
Compared to all other activities which I partake in, this one is... ()							

	Low					High	
	0	1	2	4	5	6	7
I that my competence in this area is... ()							

	Too Low			Just Right			Too high
	0	1	2	4	5	6	7

Page 10 of 16

For me personally, the current demands are...	

How engaged were you with the training?

Not at All			Somewhat			A Lot	
0	1	2	3	4	5	6	7

How immersed did you feel in the training?

Not at all			Somewhat			A Lot	
0	1	2	3	4	5	6	7

How clear were the training tasks?

Not Clear at All			Somewhat Clear			Very Clear
0	1	2	4	5	6	7

How clear was the feedback?

Not Clear at All			Somewhat Clear			Very Clear	
0	1	2	3	4	5	6	7

How appropriate was the feedback?

Not at all Appropriate	Somewhat appropriate	Very Appropriate

0	1	2	4	5	6	7

How timely was the feedback?

Not all all Timely	Somewhat Timely	Very Timely

0	1	2	4	5	6	7

Did you feel overloaded with information?

Not At All	Somewhat	Very

0	1	2	4	5	6	7

Display This Question:

If Did you feel overloaded with information? [Yes] > 4

To what extent?

Not At All	Somewhat	Very

0	1	2	4	5	6	7

How confident do you feel in performing data collection tasks by yourself?

Not at all Confident	Somewhat Confident	Very Confident

0	1	2	4	5	6	7

How motivated are you to participate in the data collection task today?

	Not at all Motivated			Somewhat Motivated			Very Motivated
	0	1	2	4	5	6	7

How motivated are you to participate in future citizen science projects?

	Not at all Motivated			Somewhat Motivated			Very Motivated
	0	1	2	4	5	6	7

Did you experience any nausea during the training?

	None		Very Little		Some		A Lot
	0	1	2	4	5	6	7

Display This Question:

If Did you experience any nausea during the training? [0] > 0

How severe was it?

	Not at all Severe				Somewhat Severe			Very Severe
	0	1	2	3	4	5	6	7

Display This Question:

If Did you experience any nausea during the training? [0] > 0

How much did it interfere with training?

	Didn't interfere at all		Somewhat interfered		Significantly interfered	
0	1	2	4	5	6	7

Display This Question:

If Did you experience any nausea during the training? [0] > 0

How many times did you have to stop?

- 0 (1)
- 1 (2)
- 2 (3)
- 3+ (4)

Additional comments:

____

(Post-Study)

How motivated were you to participate in the data collection task today?

	Not at all motivated		Somewhat motivated		Very motivated	
1	2	3	4	5	6	7

How much did you enjoy doing outdoor data collection?

Not all All		Somewhat		A Lot		
1	2	3	4	5	6	7

How engaged were you in the data collection tasks?

Not at All		Somewhat		A Lot		
1	2	3	4	5	6	7

How immersive did you find the tasks?

Not at All		Somewhat		A Lot		
1	2	3	4	5	6	7

How appropriate was the training to the data collection tasks?

Inappropriate		Neutral		Appropriate		
1	2	3	4	5	6	7

How confident do you feel in performing data collection tasks by yourself?

Not at all Confident			Somewhat Confident			Very Confident
1	2	3	4	5	6	7

How motivated are you to participate in future citizen science projects?

None			Some			A Lot
1	2	3	4	5	6	7

How likely are you to participate in future data collection projects?

Not at All			Somewhat			A Lot	
1	2	2	3	4	5	5	6

Additional Comments:

____

A.3 Multisensory Study Questions

StreamBED Video Study

Video 1

Please list different features of the environment that you saw or physically experienced. For each item, give as much detail as possible about it.

Examples:
sandy beach
fog
clear day
slow moving
water
fast moving water
a flock of birds,
sounds like 3 species
crickets tall trees,
a few feet from stream
shorter water plants in front
no grass
clear skies
one log in the water
Murky water, covered in vegetation
big tree on right side with some erosion
big stream area
can really hear water
several logs in the water
closest side covered in rocks
narrow stream on other bank, very densely covered in fern plants,
with tall trees
2 streams intersecting

As you explore the environment in Video 1, please describe different features of the environment that you **saw or physically experienced.** For each item, give as much **detail** as possible about it.

____

____

____

____

____

Page 1 of 25

How many **senses** did you experience? For each sense you selected, what **stimuli** did you notice most? (e.g. hearing - the sound of bees, sight - seeing how fast the water was moving)

- ☐ Sight (vision) (1) ____________________
- ☐ Hearing (audition) (2) ____________________
- ☐ touch (haptic or somatosensation) (3) ____________________
- ☐ smell (olfaction) (5) ____________________
- ☐ Taste (gustation) (4) ____________________

Which sense did you notice most?

- ○ Sight (vision) (1)
- ○ Hearing (audition) (2)
- ○ touch (haptic or somatosensation) (3)
- ○ smell (olfaction) (5)
- ○ Taste (gustation) (4)

To what extend did you **feel like you were surrounded** by the environment in video 1?

	Not at all					Very	
	1	2	3	4	5	6	7

How **realistic** did the environment in Video 1 feel?

	Not at all					Very	
	1	2	3	4	5	6	7

Page 2 of 25

How **familiar** did the Video 1 environment feel?

	Not at all familiar					Very familiar
1	2	3	4	5	6	7

What inferences can you draw about about the environmental habitats in Video 1?
(e.g. hearing bees in a crop field might indicate the presence of a hive nearby, healthy crops, and a lack of pesticides.)

What do you think this stream looked like in the past? How do you think it will change in the future?

Video 2

Please list different features of the environment that you s*aw or physically experienced*. For each item, give as much detail as possible about it.

Examples:
sandy beach
fog
clear day
slow moving
water
fast moving water
a flock of birds,
sounds like 3 species
crickets tall trees,
a few feet from stream
shorter water plants in front
no grass
clear skies
one log in the water
Murky water, covered in vegetation
big tree on right side with some erosion
big stream area
can really hear water
several logs in the water
closest side covered in rocks
narrow stream on other bank, very densely covered in fern plants,
with tall trees
2 streams intersecting

As you explore the environment in Video 2, please describe different features of the environment that you **saw or physically experienced.** For each item, give as much **detail** as possible about it.

____

____

____

____

____

How many **senses** did you experience? For each sense you selected, what **stimuli** did you notice most? (e.g. hearing - the sound of bees, sight - seeing how fast the water was moving)

- ☐ Sight (vision) (1) ____________________
- ☐ Hearing (audition) (2) ____________________
- ☐ touch (haptic or somatosensation) (3) ____________________
- ☐ smell (olfaction) (5) ____________________
- ☐ Taste (gustation) (4) ____________________

Which sense did you notice most?

- ○ Sight (vision) (1)
- ○ Hearing (audition) (2)
- ○ touch (haptic or somatosensation) (3)
- ○ smell (olfaction) (5)
- ○ Taste (gustation) (4)

To what extend did you **feel like you were surrounded** by the environment in Video 2?

Not at all — Very

	1	2	3	4	5	6	7

Q162 How **realistic** did the environment in Video 2 feel?

Not at all — Very

1	2	3	4	5	6	7

How **familiar** did the Video 2 environment feel?

	Not at all familiar				Very familiar	
1	2	3	4	5	6	7

What inferences can you draw about about the the environmental habitats in Video 2?
(e.g. hearing bees in a crop field might indicate the presence of a hive nearby, healthy crops, and a lack of pesticides.)

What do you think this stream looked like in the past? How do you think it will change in the future?

Video 1

Video 2

What **similarities** did you observe between habitats in Video 1 and Video 2?
(e.g. both habitats had dead grass)

____

____

____

____

____

What **differences** did you observe between habitats in Video 1 and Video 2?
(e.g. one habitat was cold and windy, while the other habitat was hot and humid)

____

____

____

____

____

What **inferences** can you make about the habitats in Videos 1 vs. 2?
(the habitat in video 1 and two are from a similar location)

____

____

____

____

____

Answer the following questions to the best of your abilities based on your experience with the **last 2 videos**

	Not at all		Somewhat		Very Much	
	0	1	3	4	6	7
How much do you expect to enjoy collecting data about streams? ()						
How excited are you to interact with future training? ()						
How much do you expect to enjoy future training about streams? ()						
How excited are you to interact with a Virtual Reality headset? ()						

Additional Comments:

____

Video 3

Please list different features of the environment that you s*aw or physically experienced*. For each item, give as much detail as possible about it.

Examples:

sandy beach
fog
clear day
slow moving
water
fast moving water
a flock of birds,
sounds like 3 species
crickets tall trees,
a few feet from stream
shorter water plants in front
no grass
clear skies
one log in the water
Murky water, covered in vegetation
big tree on right side with some erosion
big stream area
can really hear water
several logs in the water
closest side covered in rocks
narrow stream on other bank, very densely covered in fern plants,
with tall trees
2 streams intersecting

As you explore the environment in Video 3, please describe different features of the environment that you **saw or physically experienced.** For each item, give as much **detail** as possible about it.

____

____

____

____

____

How many **senses** did you experience? For each sense you selected, what **stimuli** did you notice most? (e.g. hearing - the sound of bees, sight - seeing how fast the water was moving)

- ☐ Sight (vision) (1) ____________________
- ☐ Hearing (audition) (2) ____________________
- ☐ touch (haptic or somatosensation) (3) ____________________
- ☐ smell (olfaction) (5) ____________________
- ☐ Taste (gustation) (4) ____________________

Which sense did you notice most?

- ○ Sight (vision) (1)
- ○ Hearing (audition) (2)
- ○ touch (haptic or somatosensation) (3)
- ○ smell (olfaction) (5)
- ○ Taste (gustation) (4)

To what extend did you **feel like you were surrounded** by the environment in Video 3?

	Not at all					Very	
	1	2	3	4	5	6	7

How **realistic** did the environment in Video 3 feel?

	Not at all					Very	
	1	2	3	4	5	6	7

How **familiar** did the Video 3 environment feel?

	Not at all familiar						Very familiar
	1	2	3	4	5	6	7
1 ()							

What inferences can you draw about about the environmental habitats in Video 3?
(e.g. hearing bees in a crop field might indicate the presence of a hive nearby, healthy crops, and a lack of pesticides.)

____

____

____

____

____

What do you think this stream looked like in the past? How do you think it will change in the future?

____

____

____

____

____

Video 4

Please list different features of the environment that you s*aw or physically experienced*. For each item, give as much detail as possible about it.

Examples:

sandy beach
fog
clear day
slow moving
water
fast moving water
a flock of birds,
sounds like 3 species
crickets tall trees,
a few feet from stream
shorter water plants in front
no grass
clear skies
one log in the water
Murky water, covered in vegetation
big tree on right side with some erosion
big stream area
can really hear water
several logs in the water
closest side covered in rocks
narrow stream on other bank, very densely covered in fern plants,
with tall trees
2 streams intersecting

As you explore the environment in Video 4, please describe different features of the environment that you **saw or physically experienced.** For each item, give as much **detail** as possible about it.

____

____

____

____

____

How many **senses** did you experience? For each sense you selected, what **stimuli** did you notice most? (e.g. hearing - the sound of bees, sight - seeing how fast the water was moving)

- ☐ Sight (vision) (1) ____________________
- ☐ Hearing (audition) (2) ____________________
- ☐ touch (haptic or somatosensation) (3) ____________________
- ☐ smell (olfaction) (5) ____________________
- ☐ Taste (gustation) (4) ____________________

Which sense did you notice most?

- ○ Sight (vision) (1)
- ○ Hearing (audition) (2)
- ○ touch (haptic or somatosensation) (3)
- ○ smell (olfaction) (5)
- ○ Taste (gustation) (4)

To what extend did you **feel like you were surrounded** by the environment in video 4?

Not at all Very

1 2 3 4 5 6 7

How **realistic** did the environment in Video 4 feel?

Not at all Very

1 2 3 4 5 6 7

How **familiar** did the Video 4 environment feel?

	Not at all familiar				Very familiar	
1	2	3	4	5	6	7

What inferences can you draw about the environmental habitats in Video 4?
(e.g. hearing bees in a crop field might indicate the presence of a hive nearby, healthy crops, and a lack of pesticides.)

What do you think this stream looked like in the past? How do you think it will change in the future?

Video 3

Video 4

What **similarities** did you observe between habitats in Video 3 and Video 4?
(e.g. both habitats had dead grass)

____

____

____

____

What **differences** did you observe between habitats in Video 3 and Video 4?
(e.g. one habitat was cold and windy, while the other habitat was hot and humid)

What **inferences** can you make about the habitats in Video 3 and Video 4?
(the habitat in video 1 and two are from a similar location)

Answer the following questions to the best of your abilities based on your experience with the **last 2 videos**

	Not at all		Somewhat		Very Much	
	0	1	3	4	6	7
How much do you expect to enjoy collecting data about streams? ()						
How excited are you to interact with future training? ()						
How much do you expect to enjoy future training about streams? ()						
How excited are you to interact with a Virtual Reality headset? ()						

Additional Comments:

Which set of videos was most immersive and realistic?

- ○ Videos 1 & 2 (1)
- ○ Videos 3 & 4 (2)

Display This Question:

If Which set of videos was most immersive and realistic? = Videos 1 & 2

Why were Videos 1 & 2 more realistic than 3 & 4?

____

____

____

____

____

Display This Question:

If Which set of videos was most immersive and realistic? = Videos 3 & 4

Why were Videos 3 & 4 more realistic than Videos 1 & 2?

____

____

____

____

____

Which set of videos did you prefer?

- ○ Videos 1 & 2 (1)
- ○ Videos 3 & 4 (2)

Which set of videos best kept your attention?

- ◯ Videos 1 & 2 (1)
- ◯ Videos 3 & 4 (2)

Which set of videos were the most interesting?

- ◯ Videos 1 & 2 (1)
- ◯ Videos 3 & 4 (2)

How many citizen science projects are you familiar with?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

Display This Question:

If How many citizen science projects are you familiar with? != 0

How many citizen science projects have you been involved in?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

Display This Question:

If How many citizen science projects are you familiar with? != 0

Did you participate in data collection on a citizen science project?

- ◯ Yes (1)
- ◯ No (2)

Display This Question:
If Did you participate in data collection on a citizen science project? = Yes

If so, for the project you were most involved in:

Display This Question:
If so, for the project you were most involved in: Is Displayed

How many times did you participate in data collection?

- ◯ 1 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (4)
- ◯ 5+ (3)

Display This Question:
If so, for the project you were most involved in: Is Displayed

How comfortable do you feel collecting data?

	Not Comfortable at all		Somewhat Comfortable		Very Comfortable	
0	1	2	4	5	6	7

Display This Question:
If so, for the project you were most involved in: Is Displayed

Q5 What types of training did you receive?

- ☐ Pamphlets (1)
- ☐ Online resources (2)
- ☐ Lectures (3)
- ☐ Game (4)
- ☐ On-site training (5)

Display This Question:

If so, for the project you were most involved in: Is Displayed

How many hours of training did you receive?

- ○ 0 (1)
- ○ 1-2 (2)
- ○ 3-4 (3)
- ○ 5+ (4)

How familiar are you with water quality monitoring?

	Not Familiar at All			Somewhat Familiar			Very Familiar
	1	2	3	4	5	6	7

Display This Question:

How familiar are you with water quality monitoring? [5] > 1

Have you ever participated in water quality monitoring data collection?

- ○ Yes (1)
- ○ No (2)

Display This Question:

Have you ever participated in water quality monitoring data collection? = Yes

How many projects did you participate in?

- ◯ 1 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (4)
- ◯ 5+ (3)

Display This Question:

Have you ever participated in water quality monitoring data collection? = Yes

For the project you were most involved in:

Display This Question:

Have you ever participated in water quality monitoring data collection? = Yes

How many hours of training did you receive?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

Display This Question:

If so, for the project you were most involved in: Is Displayed

How comfortable do you feel collecting data?

	Not Comfortable at all		Somewhat Comfortable		Very Comfortable	
0	1	2	4	5	6	7

()

Display This Question:

Have you ever participated in water quality monitoring data collection? = Yes

How many times did you participate?

- ◯ 1 (1)
- ◯ 2-3 (4)
- ◯ 4-5 (5)
- ◯ 6+ (6)

[answer the following questions to the best of your abilities]

	Never		Sometimes		Always		
	0	1	2	4	5	6	7
I can understand all of the key concepts covered in my course. ()							
Soon after the end of a lesson, I am able to remember all of the key concepts ()							
Soon after the end of a lesson, I am always able to distinguish the most important concepts from concepts of less importance ()							
It is always easy for me to understand new information, even on a topic that does not interest me very much ()							
It is always easy for me to connect new information about a topic that interests me with other pieces of information ()							
When I find something new about a topic that I am studying, I am always able to connect it with other things that I know about the topic ()							
During a course, if we are given a new task to complete, I can always complete it by applying the knowledge that I obtained from lessons ()							

How many hours do you spend playing computer-based video games per week?

- ○ 0 (1)
- ○ 1-2 (5)
- ○ 3-4 (6)
- ○ 5+ (2)

Have you ever experienced virtual reality?

- ○ Yes (1)
- ○ No (2)

Display This Question:

Have you ever experienced virtual reality? = Yes

How comfortable did you feel interacting in a virtual reality environment?

	Not comfortable at all			Somewhat comfortable			Very Comfortable	
	0	1	2	3	4	5	6	7

Are you...

- ○ Male (1)
- ○ Female (2)

Which category best describes your ethnicity?

- ◯ African American (4)
- ◯ Asian American (5)
- ◯ Hispanic (6)
- ◯ Native American (7)
- ◯ White/Caucasian (8)
- ◯ Other (9)

What's the highest level of education you've completed?

- ◯ High School (1)
- ◯ Bachelors Degree (2)
- ◯ Masters degree (3)
- ◯ PhD (4)

Page 25 of 25

A.4 Final Study VR Questions

StreamBED Final Study

Participant Number

____

Partner

A

B

Condition

A

B

VR [A] - Taking Photos

To what extend did you **feel like you were surrounded** by the environment in this stream?

	Not at all				Very	
1	2	3	4	5	6	7

How **realistic** did the environment feel?

	Not at all				Very	
1	2	3	4	5	6	7

How **familiar** did the environment feel?

	Not at all familiar				Very familiar	
1	2	3	4	5	6	7

Page 1 of 23

How **engaged** were you by the VR task?

Not at all Engaged Very Engaged

1 2 3 4 5 6 7

How easy was it easy to **interact** with the VR environment?

Very easy Very difficult

1 2 3 4 5 6 7

How easy was it to **collaborate** on the VR task?

Very easy Very difficult

1 2 3 4 5 6 7

VR [A] - Partner

How **engaged** were you by the VR task?

Not at all Engaged Very Engaged

1 2 3 4 5 6 7

How easy was it to **contribute** to the VR task?

	Very easy				Very difficult	
1	2	3	4	5	6	7

Reference [A] - Assessment

Epifaunal Substrate Rating?

0 1 2 3 4 5 6 7 8 9 1011121314151617181920

How confident you are in your rating?

- Not at all Confident (4)
- (13)
- (14)
- Somewhat confident (15)
- (11)
- (12)
- Very confident (17)

How much input did **you** have in the final rating?

	Not all all				Very much	
1	2	3	4	5	6	7

Page 3 of 23

How much input did your partner have in the final rating?

	Not all all			Very much		
1	2	3	4	5	6	7

Bank Stability Rating?

0	1	2	3	4	5	6	7	8	9	10

How **confident** you are in your rating?

- ○ Not at all Confident (4)
- ○ (13)
- ○ (14)
- ○ Somewhat confident (15)
- ○ (11)
- ○ (12)
- ○ Very confident (17)

How much input did **you** have in the final rating?

	Not all all			Very much		
1	2	3	4	5	6	7

How much input did your partner have in the final rating?

	Not all all				Very much	
1	2	3	4	5	6	7

Overall, how **easy was it to interact** with the desktop environment?

	Very easy				Very difficult	
1	2	3	4	5	6	7

Reference [A] - Partner

How much input did **you** have in the final **Epifaunal Substrate r**ating?

	Not all all				Very much	
1	2	3	4	5	6	7

How much input did your partner have in the final **Epifaunal Substrate** rating?

	Not all all				Very much	
1	2	3	4	5	6	7

Page Break

How much input did **you** have in the final **Bank Stability** rating?

Not all all — Very much

1 2 3 4 5 6 7

How much input did your partner have in the final **Bank Stability** rating?

Not all all — Very much

1 2 3 4 5 6 7

Overall, How easy was it to **contribute** to desktop task?

Very easy — Very difficult

1 2 3 4 5 6 7

VR [B] - Partner

How **engaged** were you by the VR task?

Not at all Engaged — Very Engaged

1 2 3 4 5 6 7

How easy was it to contribute to the VR task?

Very easy — Very difficult

1 2 3 4 5 6 7

VR [B] - Taking Photos

To what extend did you **feel like you were surrounded** by the environment in this stream?

Not at all   Very

1 2 3 4 5 6 7

How **realistic** did the environment feel?

Not at all   Very

1 2 3 4 5 6 7

How **familiar** did the environment feel?

Not at all familiar   Very familiar

1 2 3 4 5 6 7

How **engaged** were you by the VR task?

Not at all Engaged   Very Engaged

1 2 3 4 5 6 7

How easy was it easy to interact with the VR environment?

	Very easy			Very difficult			
	1	2	3	4	5	6	7

How easy was it to **collaborate** on the VR task?

	Very easy			Very difficult			
	1	2	3	4	5	6	7

Reference [B] - Assessment

Epifaunal Substrate Rating?

0 1 2 3 4 5 6 7 8 9 1011121314151617181920

How confident you are in your rating?

- Not at all Confident (4)
- (13)
- (14)
- Somewhat confident (15)
- (11)
- (12)
- Very confident (17)

How much input did **you** have in the final rating?

	Not all all				Very much	
1	2	3	4	5	6	7

How much input did your partner have in the final rating?

	Not all all				Very much	
1	2	3	4	5	6	7

Bank Stability Rating?

0	1	2	3	4	5	6	7	8	9	10

How **confident** you are in your rating?

- ◯ Not at all Confident (4)
- ◯ (13)
- ◯ (14)
- ◯ Somewhat confident (15)
- ◯ (11)
- ◯ (12)
- ◯ Very confident (17)

How much input did **you** have in the final rating?

Not all all — Very much

1 2 3 4 5 6 7

How much input did your partner have in the final rating?

Not all all — Very much

1 2 3 4 5 6 7

How easy was it easy to **interact** with the desktop environment?

Very easy — Very difficult

1 2 3 4 5 6 7

Reference [B] - Partner

How much input did **you** have in the final **epifaunal substrate** rating?

Not all all — Very much

1 2 3 4 5 6 7

How much input did your partner have in the final **epifaunal substrate** rating?

Not all all — Very much

1 2 3 4 5 6 7

How much input did **you** have in the final rating?

Not all all				Very much		
1	2	3	4	5	6	7

How much input did your partner have in the final **Bank Stability** rating?

Not all all				Very much		
1	2	3	4	5	6	7

Overall, How easy was it to **contribute** to desktop task?

Very easy				Very difficult		
1	2	3	4	5	6	7

In the next section, you will answer questions about your **VR training experience**

How **clear** were the VR **training tasks?**

Not Clear at All			Somewhat Clear			Very Clear
1	2	3	4	5	6	7

How **clear** was the **VR feedback?**

Not Clear at All			Somewhat Clear				Very Clear
1	2	3	4	4	5	6	7

How **appropriate** was the **VR feedback**?

	Not at all Appropriate		Somewhat appropriate		Very Appropriate	
1	2	3	4	5	6	7

How **timely** was the **VR feedback**?

	Not all all Timely		Somewhat Timely		Very Timely	
1	2	3	4	5	6	7

Did you feel **overloaded** with information?

	Not At All		Somewhat		Very	
1	2	3	4	5	6	7

Display This Question:

If Did you feel overloaded with information? [1] > 4

To what extent?

	Not At All		Somewhat		Very	
1	2	3	4	5	6	7

In the next section, you will answer questions about your **Desktop Training Experience**

How **clear** were the **desktop training tasks?**

Not Clear at All		Somewhat Clear		Very Clear		
1	2	3	4	5	6	7

How **clear** was the **desktop feedback?**

Not Clear at All		Somewhat Clear		Very Clear			
1	2	3	4	4	5	6	7

How **appropriate** was the **desktop feedback**?

Not at all Appropriate		Somewhat appropriate		Very Appropriate		
1	2	3	4	5	6	7

How **timely** was the **desktop feedback**?

Not all all Timely		Somewhat Timely		Very Timely		
1	2	3	4	5	6	7

Did you feel **overloaded** with information?

	Not At All		Somewhat		Very	
1	2	3	4	5	6	7

Display This Question:

If Did you feel overloaded with information? [1] > 4

To what extent?

	Not At All		Somewhat		Very	
1	2	3	4	5	6	7

How **confident** would you feel in performing data collection tasks with your partner?

	Not at all Confident		Somewhat Confident		Very Confident	
1	2	3	4	5	6	7

In the next section, you will answer questions about outdoor data collection

To what extent did you enjoy assessing the outdoor stream?

- ◯ Very much (1)
- ◯ (5)
- ◯ (6)
- ◯ Somewhat (9)
- ◯ (10)
- ◯ (11)
- ◯ Not at all (12)

How challenging was it to assess the outdoor stream?

- ◯ Very challenging (1)
- ◯ (5)
- ◯ (6)
- ◯ Somewhat challenging (9)
- ◯ (10)
- ◯ (11)
- ◯ Not challenging at all (12)

To what extent did you rely on your partner to make the outdoor assessments?

- ◯ Not at all (1)
- ◯ (5)
- ◯ (6)
- ◯ Somewhat (9)
- ◯ (10)
- ◯ (11)
- ◯ Very much (12)

How excited are you to interact with future training?

- ◯ Very much (1)
- ◯ (5)
- ◯ (6)
- ◯ Somewhat (9)
- ◯ (10)
- ◯ (11)
- ◯ Not at all (12)

How much do you expect to enjoy collecting data about streams?

- Not at all (1)
- (5)
- (6)
- Somewhat (9)
- (10)
- (11)
- Very much (12)

In the next section, you will answer questions about your experience with citizen science and water monitoring

How many citizen science projects are you familiar with?

- 0 (1)
- 1-2 (2)
- 3-4 (3)
- 5+ (4)

Display This Question:

If How many citizen science projects are you familiar with? != 0

How many citizen science projects have you been involved in?

- 0 (1)
- 1-2 (2)
- 3-4 (3)
- 5+ (4)

Display This Question:
If How many citizen science projects are you familiar with? != 0

Did you participate in data collection on a citizen science project?

- ◯ Yes (1)
- ◯ No (2)

Display This Question:
If Did you participate in data collection on a citizen science project? = Yes

If so, for the project you were most involved in:

Display This Question:
If so, for the project you were most involved in: Is Displayed

How many times did you participate in data collection?

- ◯ 1 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (4)
- ◯ 5+ (3)

Display This Question:
If so, for the project you were most involved in: Is Displayed

How comfortable do you feel collecting data?

Not Comfortable at all	Somewhat Comfortable	Very Comfortable

0	1	2	4	5	6	7

Display This Question:

If so, for the project you were most involved in: Is Displayed

What types of training did you receive?

- ☐ Pamphlets (1)
- ☐ Online resources (2)
- ☐ Lectures (3)
- ☐ Game (4)
- ☐ On-site training (5)

Display This Question:

If so, for the project you were most involved in: Is Displayed

How many hours of training did you receive?

- ○ 0 (1)
- ○ 1-2 (2)
- ○ 3-4 (3)
- ○ 5+ (4)

How familiar are you with water quality monitoring?

	Not Familiar at All			Somewhat Familiar			Very Familiar
	1	2	3	4	5	6	7

Display This Question:

If How familiar are you with water quality monitoring? [5] > 1

Have you ever participated in water quality monitoring data collection?

- ◯ Yes (1)
- ◯ No (2)

Display This Question:

If Have you ever participated in water quality monitoring data collection? = Yes

How many projects did you participate in?

- ◯ 1 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (4)
- ◯ 5+ (3)

Display This Question:

If Have you ever participated in water quality monitoring data collection? = Yes

For the project you were most involved in:

Display This Question:

If Have you ever participated in water quality monitoring data collection? = Yes

How many hours of training did you receive?

- ◯ 0 (1)
- ◯ 1-2 (2)
- ◯ 3-4 (3)
- ◯ 5+ (4)

Display This Question:

If If so, for the project you were most involved in: Is Displayed

How comfortable do you feel collecting data?

	Not Comfortable at all		Somewhat Comfortable		Very Comfortable	
0	1	2	4	5	6	7

Display This Question:

If Have you ever participated in water quality monitoring data collection? = Yes

How many times did you participate?

- 1 (1)
- 2-3 (4)
- 4-5 (5)
- 6+ (6)

In the next section, you will answer questions about your general learning experience and background.

How many hours do you spend playing computer-based video games per week?

- 0 (1)
- 1-2 (5)
- 3-4 (6)
- 5+ (2)

Have you ever experienced virtual reality?

◯ Yes (1)

◯ No (2)

Display This Question:

If Have you ever experienced virtual reality? = Yes

How comfortable did you feel interacting in a virtual reality environment?

	Not comfortable at all			Somewhat comfortable			Very Comfortable	
	0	1	2	3	4	5	6	7

Are you...

◯ Male (1)

◯ Female (2)

Which category best describes your ethnicity?

◯ African American (4)

◯ Asian (5)

◯ Hispanic (6)

◯ Native American (7)

◯ White/Caucasian (8)

◯ Other (9) ____________________

Page 22 of 23

What age category best describes you?

- ◯ 18-24 (1)
- ◯ 25-34 (4)
- ◯ 35-44 (5)
- ◯ 45-54 (6)
- ◯ 55-64 (7)
- ◯ 65+ (8)

What's the highest level of education you've completed?

- ◯ High School (1)
- ◯ Bachelors Degree (2)
- ◯ Masters degree (3)
- ◯ PhD (4)

How well did you know your partner before the study?

- ◯ Not at all (Never met them before) (1)
- ◯ Barely (2)
- ◯ Somewhat (3)
- ◯ Somewhat Well (4)
- ◯ Very Well (We're close) (5)

A.5 Final Study Discussion Questions

A.5.1 After Training

1. What did you think of the VR and desktop training?

2. Was there anything you liked or disliked about training?

3. Did you notice the ambient multisensory experience? Did you think it affected your training experience? (Multisensory VR Condition)

4. How did it feel working with a partner that you [knew /did not know]?

5. Do you think it would have easier or more difficult to complete the training on your own? Why?

6. The Maryland stream you experienced was meant to be an initial training experience; the idea is that over time, you would have the chance to experience more streams. Would you want to continue working with a partner, or would you want to complete future training on your own? Why or why not?

A.5.2 After Outdoor Evaluation

1. What did you think of the Outdoor Evaluation? How difficult was it?

2. What steps did you go through during the outdoor evaluation? Which steps did you complete with your partner vs. by yourself?

3. Having completed the outdoor evaluation, what did you think of the VR and desktop training? Was there anything you liked or disliked?

4. Having completed the outdoor evaluation, do you think the multisensory training experience affected your training? Why or why not? (Multisensory VR Condition)

A.6 Final Study Powerpoint Questions

Overall, how **realistic** was the PPT training?

1. Very realistic
2.
3. Somewhat realistic
4.
5. Not at all realistic

Overall, how **familiar** did the PPT training environment feel?

6. Very familiar
7.
8. Somewhat familiar
9.
10. Not at all familiar

Overall, how **easy was it to contribute** to the PPT training discussion?

1. Not at all easy
2.
3. Somewhat easy
4.
5. Very Easy

Overall, how **engaged** were you by the Powerpoint (PPT) Training?

1. Not engaged at all
2.
3. Somewhat Engaged
4.
5. Very Engaged

Overall, how **motivated** were you by the Powerpoint (PPT) Training?

1. Very motivated
2.
3. Somewhat motivated
4.
5. Not at all motivated

To what extent are you **looking forward to** performing outdoor data collection tasks?

1. Not at all
2.
3. Somewhat
4.
5. Very

Answer the following questions to the best of your ability:

To what extent did you enjoy assessing the outdoor stream?

Very much			Somewhat			Not at all

How challenging was it to assess the outdoor stream?

Very challenging			Somewhat challenging			Not challenging at all

To what extent did you rely on your partner to make the outdoor assessments?

Not at all			Somewhat			Very much

How excited are you to interact with future training?

Very much			Somewhat			Not at all

How much do you expect to enjoy collecting data about streams?

Not at all			Somewhat			Very much

Citizen Science Background:

How many citizen science projects are you familiar with?

1. 0
2. 1-2
3. 3-4
4. 5+

How many citizen science projects have you been involved in?

1. 0
2. 1-2
3. 3-4
4. 5+

Did you participate in data collection on a citizen science project?

1. No
2. Yes

Familiarity with Water Quality Monitoring:

How many water water quality projects have you participated in?

1. 0
2. 1-2
3. 3-4
4. 5+

If you participated in a project, for the project you were most involved in, how many hours of training did your recieve?

1. 0
2. 1-2
3. 3-4
4. 5+

How many times did you participate in water quality data collection?

1. 0
2. 1-2
3. 3-4
4. 5+

Are you….

1. Male
2. Female
3. Prefer not to say

What category best describes your ethnicity:

1. African American
2. Asian
3. Hispanic
4. Native American
5. White/Caucasian
6. Other ________________

What age category best describes you?

1. 18-24
2. 25-34
3. 35-44
4. 54-54
5. 55-64
6. 65+

What is the highest level of education you've completed?

1. High School
2. Some College
3. Bachelor's Degree
4. Masters Degree
5. PhD

How well did you know your partner before the study?

1. Not at all (Never met them before)
2. Barely
3. Somewhat
4. Somewhat Well
5. Very Well (We're close)

A.7 Final Study Training Measures and Correlations

VR Training Engagement

Measures
1. VR Training Immersion 2. VR Training Realism 3. VR Environment Familiarity 4. VR Engagement 5. VR Partner's Engagement

Correlations Between Measures
VR Realism /VR Familiarity (ρ=.707)
VR Realism / VR Engagement (ρ=.612)

Familiarity/Immersion (ρ=.317)

Partner Engagement / Immersion (ρ=.446)
Partner Engagement / Familiarity (ρ=.336)
Partner Engagement / VR Engagement (ρ=.188)

PPT Training Engagement

Measures
1. PPT Training Realism 2. PPT Training Familiarity 3. PPT Training Engagement 4. Motivation to collect data as a result of PPT training 5. Looking forward to Outdoor Assessment

Correlations Between Measures
VR Realism/ Looking Forward to Data Collection = (ρ=-.29)
Engagement / Looking Forward to Data Collection(ρ=.62)

Motivation / Looking Forward to Data Collection(ρ=.50)
Motivation / Engagement (ρ=.38)

VR Training Interaction

VR Interaction
1. Ease of interacting with VR system 2. Ease of collaborating using VR system 3. Partner's feeling of contributing to the VR experience 4. VR feedback clarity 5. VR feedback appropriateness 6. VR feedback timeliness 7. VR system information overload

<u>Correlations Between Measures</u>
Ability to Interact / VR training Clarity (ρ=.653)
Ability to Interact / VR Feedback Clarity (ρ=.780)
Ability to Interact / VR Feedback Timeliness (ρ=.769),
Ability to Interact / VR Feedback Appropriateness (ρ=.970)
Ability to Interact / VR Participant Overload (ρ=.405)

Collaborate with Partner / Partner's ability to Contribute to VR (ρ=.944)
Collaborate with Partner / VR Clarity (ρ=.911)
Collaborate with Partner / VR Feedback Clarity (ρ=.784)
Collaborate with Partner / VR Feedback Appropriateness (ρ=.596)
Collaborate with Partner / VR Participant Overload (ρ=.430)

Partner's Ability to Contribute / VR Clarity (ρ=.822)
Partner's Ability to Contribute / VR Feedback Timeliness (ρ=.737
Partner's Ability to Contribute / VR Feedback Appropriateness (ρ=.530)

VR Overloaded / VR Feedback Timeliness (ρ=.295).

Desktop Training Interaction

Desktop Interaction
1. Ease of Interaction with Desktop system 2. Ease of Collaborating Using Desktop System 3. Partner's Contribution to Desktop Experience 4. Desktop Feedback Clarity 5. Desktop Feedback Appropriateness 6. Desktop Feedback Timeliness 7. Desktop System Information Overload

<u>Correlations Between Measures</u>
Desktop interaction / Partner's Ability to Contribute (ρ=-0.56)
Desktop interaction / Desktop Participant Overload (ρ=0.35)
Desktop interaction / Desktop Feedback Clarity (ρ=-0.37)
Desktop interaction / Desktop Feedback Appropriateness (ρ=-0.27)
Desktop interaction / Desktop Feedback Timeliness (ρ=-0.47)

Desktop Clarity / Desktop Feedback Clarity (ρ=0.76)
Desktop Clarity / Desktop Feedback Appropriateness (ρ=-0.54)
Desktop Clarity / Feedback Timeliness (ρ=0.37),
Desktop Clarity / Desktop Participant Overload (ρ=-.60)
Desktop Clarity / Partner's Ability to Contribute (ρ=-0.41)

Partner's Ability to Contribute / Desktop Feedback Appropriateness (ρ=.81)
Partner's Ability to Contribute / Desktop Feedback Timeliness (ρ=.44)
Partner's Ability to Contribute / Desktop Feedback Clarity (ρ=-0.27)
Partner's Ability to Contribute / Desktop Feedback Appropriateness (ρ=-0.39)
Partner's Ability to Contribute / Desktop Feedback Timeliness (ρ=-0.49)
Partner's Ability to Contribute / Participant Overload (ρ=-.56)

Desktop Feedback Appropriateness/ Desktop Feedback Timeliness (ρ=.65)
Desktop Feedback Appropriateness/ Desktop Participant Overload (ρ=-.48)

PPT Training Interaction

PPT Interaction
1. Ability to Contribute to Training

Outdoor Engagement and Interaction (All Conditions)

Outdoor engagement metrics include 1) outdoor assessment enjoyment 2) outdoor assessment challenge, 3) participant reliance on their partner 4) excitement for future training, and excitement for future data collection. In addition to comparing engagement, we compared final participant scores across the three conditions (1) Epifaunal Substrate, 2) Left Bank Stability, and 3) Right Bank Stability) relatively to an expert ``gold standard.''

Measures
1) Outdoor Assessment Enjoyment 2) Outdoor Assessment Challenge 3) Participant Reliance on Partner during Assessment 4) Excitement for Future Training 5) Excitement for Future Data Collection.

Outdoor Assessment Scores

(Absolute Difference between Participant and Expert Scores)

Measures
1. Epifaunal Substrate 2. Left Bank Stability 3. Right Bank Stability

A.8 Final Study Codebook

Final Study Initial Codebook

Theme	Code/Question	*Positive*	*Negative*
Realism	**VR Realism -** *how visually realistic/natural is the VR experience? #VRRealisim*	*"I felt like I was in the training world"*	*" It was hard to tell where different things in the environment were"*
	Desktop image realism - *how realistic are the reference images? #DesktopRealism*	*" I could identify with the reference image"*	*"The desktop images didn't feel natural"*
	Training match to Real World - *How well does training match the real world? #TrainingMatch* *- Materials or interaction*	*" The outdoor assessment was really similar to the training"* *"It was easy to transition from training to the real world"*	*" I navigated the training environment very differently than outside"*
	Effect of Multisensory cues *Did participants notice them? Did they like being exposed to them? How effective were they? #MultisensoryCues*	*"The multisensory cues felt realistic and helped me make sense of the stream space"*	*" I didn't really pay attention to the cues...they were unnecessary"*
Colearning	**Familiarity/comfort with Partner** - *how comfortable were participants with their partner? Difference in familiarity? #PartnerComfort*	*"My partner and I were comfortable asking each other questions "*	*" I didn't feel comfortable disagreeing with my partner because we're friends…[or because we don't know each other]"*
	Negotiation of meaning/Negotiating Qualitative Judgments -	*"As we navigated the VR experience, I*	*"It was hard to talk about what we were seeing*

	how did partners make sense of the VR and Desktop and training together? #PartnerNegotiationMeaning - *Also applies to outdoor* *(okay if counterfactual)*	*asked my partner whether they were seeing what I was seeing"*	*because we were on different displays"*
	Decision making with a partner - *How did they make decisions about what to do during training? #PartnerDecisionMaking*	*" I checked my answer against my partner's, and we discussed why we chose our anwer"*	*" we argued a lot over the correct answer, and it was really frustrating"*
	Preference/Enjoyment of working with a partner vs alone - *How much did partners enjoy the experience of working together? Did they have a preference for working alone or with someone in future? #EnjoymentPreferencePartner*	*" I enjoyed comparing my answer against my partner's"* *"I found it useful to have the partner there"*	*" I think I would have made a faster or better set of decisions alone"*
Scaffolding Learning	**Synthesizing Observations into larger Patterns** - how did participants make sense of the observations? what patterns did participants notice? Were they able to form some sort of larger	*" i started seeing the relationship of the different bank features"*	*"it was hard for me to compare the training features to the outdoor assessment features"*

	structure around patterns? #TrainingPatterns - Explicitly trying to see a pattern and connect it to the assessment -- - Care more about comparing features		
	developing intuition - did participants develop intuition when making judgments? #Intuition (self-assessment) - Intuition for procedural task - Making outdoor assessment - Making assessment - Identifying things in the environment *not -- recognizing whether to start training task, not intuition for procedure	*" after the training, I had a good feeling of what the stream assessment should be"*	*"Because the reference images were different than the actual stream, I don't feel like I have a good sense of the stream ratings"*
	Environment storytelling - Did participants rely on or create narratives during the study? #EnvironmentStorytelling	*"I started thinking what could have caused the stream to look this way….and how it might look like after a flood"*	*" it was hard to understand how the stream had come to look this way"*

Overall Training Interaction	**Training feedback** - *Did participants receive adequate feedback during the training? #TrainingFeedback* - *How participants used or reacted to feedback*	*" I felt like the training gave me enough feedback on my assessments*	*"It would have helped to have gotten more detailed feedback about the expert assessment"*
	Training Clarity/interaction - what made sense? What didn't #TrainingInteraction - Just interaction, not materials - Talking about how easy it was to do the assessment, because of the training -	*" the training was really straightforward"* *"Becaues of the structured training, I knew where to start the outdoor assessment"*	*"The training felt overly complicated"* *"I wanted a pointer in the VR that could help my partner see what I was seeing"*
	Protocol Clarity - *Did participants have trouble understanding the protocol? #ProtocolClarity*	*"The protocol made sense"*	*"The protocol was hard to understand"*
Enjoyment and Interest	**Training enjoyment** - *Did participants enjoy training? #TrainingEnjoyment*	*" I enjoyed doing the training"* *"the VR was so cool"*	*" I found the training boring and too long"*
	interest in monitoring and citizen science - *did training affect their interest in citizen science or*	*"The training made me think more about streams and*	*" I don't think this training will make want to understand my habitat"*

	streams? Were they engaged by learning something new? #CitizenScienceWaterMonitoringInterest	*ecology...and I might notice streams more in my area"*	

Bibliography

[1] Handbook of biological statistics. URL http://www.biostathandbook.com/kruskalwallis.html.

[2] Sun Joo Grace Ahn, Jeremy N Bailenson, and Dooyeon Park. Short-and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior. *Computers in Human Behavior*, 39:235–245, 2014.

[3] Eugenio Alberdi, Derek H Sleeman, and Meg Korpi. Accommodating surprise in taxonomic tasks: The role of expertise. *Cognitive Science*, 24(1):53–91, 2000.

[4] Jonathan Z. Bakdash, Jason S. Augustyn, and Dennis R. Proffitt. Large displays enhance spatial knowledge of a virtual environment. In *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization*, APGV '06, pages 59–62, New York, NY, USA, 2006. ACM. ISBN 1-59593-429-4. doi: 10.1145/1140491.1140503. URL http://doi.acm.org/10.1145/1140491.1140503.

[5] Albert Bandura. Self-regulation of motivation and action through goal systems. In *Cognitive perspectives on emotion and motivation*, pages 37–61. Springer, 1988.

[6] Michael T Barbour, Jeroen Gerritsen, Blaine D Snyder, and James B Stribling. Rapid bioassessment protocols for use in wadeable streams and rivers. *Periphyton, Benthic Macroinvertebrates, and Fish (2nd edn). US Environmental Protection Agency, Office of Water, Washington, DC EPA*, 1999.

[7] Tomasz Bednarz, June Kim, Ross Brown, Allan James, Kevin Burrage, Sam Clifford, Jacqueline Davis, Kerrie Mengersen, Erin Peterson, Steven Psaltis, et al. Virtual reality for conservation. In *Proceedings of the 21st International Conference on Web3D Technology*, pages 177–178. ACM, 2016.

[8] H John B Birks, Oliver Heiri, Heiki Seppä, and Anne E Bjune. Strengths and weaknesses of quantitative climate reconstructions based on late-quaternary biological proxies. *Open Ecology Journal*, 3(1):68–110, 2011.

[9] Rick Bonney, Caren B Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V Rosenberg, and Jennifer Shirk. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11): 977–984, 2009.

[10] Christos Bouras, Alexandros Philopoulos, and Th Tsiatsos. e-learning through distributed virtual environments. *Journal of Network and Computer Applications*, 24(3):175–199, 2001.

[11] Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. Using gamification to inspire new citizen science volunteers. In *Proceedings of the first international conference on gameful design, research, and applications*, pages 18–25. ACM, 2013.

[12] Anne Elizabeth Bowser. *Cooperative design, cooperative science: Investigating collaborative research through design with floracaching.* PhD thesis, University of Maryland, College Park, 2016.

[13] Frederick P Brooks Jr, Ming Ouh-Young, James J Batter, and P Jerome Kilpatrick. Project gropehaptic displays for scientific visualization. In *ACM SIGGraph computer graphics*, volume 24, pages 177–185. ACM, 1990.

[14] C Bryanton, J Bosse, Marie Brien, Jennifer Mclean, Anna McCormick, and Heidi Sveistrup. Feasibility, motivation, and selective motor control: virtual reality compared to conventional home exercise in children with cerebral palsy. *Cyberpsychology & behavior*, 9(2):123–128, 2006.

[15] Alex Butler, Otmar Hilliges, Shahram Izadi, Steve Hodges, David Molyneaux, David Kim, and Danny Kong. Vermeer: direct interaction with a 360 viewable 3d display. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 569–576. ACM, 2011.

[16] F. Buttussi and L. Chittaro. Effects of different types of virtual reality display on presence and learning in a safety training scenario. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1063–1076, Feb 2018. ISSN 1077-2626. doi: 10.1109/TVCG.2017.2653117.

[17] Victor Candas and André Dufour. Thermal comfort: multisensory interactions? *Journal of physiological anthropology and applied human science*, 24 (1):33–36, 2005.

[18] Jeffrey P Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197, 2008.

[19] Alexandra Covaci, Gheorghita Ghinea, Chang-Hsin Lin, Shu-Hsien Huang, and Ju-Ling Shih. Multisensory games-based learning-lessons learnt from olfactory enhancement of a digital board game. *Multimedia Tools and Applications*, pages 1–19, 2018.

[20] Mary M Crossan, Henry W Lane, and Roderick E White. An organizational learning framework: From intuition to institution. *Academy of management review*, 24(3):522–537, 1999.

[21] Elaine M Crowder. Gestures at work in sense-making science talk. *The Journal of the Learning Sciences*, 5(3):173–208, 1996.

[22] Kevin Crowston. Gravity spy: Humans, machines and the future of citizen science. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, pages 163–166, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4688-7. doi: 10.1145/3022198.3026329. URL `http://doi.acm.org/10.1145/3022198.3026329`.

[23] James J Cummings and Jeremy N Bailenson. How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, 19(2):272–309, 2016.

[24] Jason Day and Jim Foley. Evaluating web lectures: A case study from hci. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pages 195–200. ACM, 2006.

[25] Jason Allan Day, James D Foley, and Richard Catrambone. Investigating multimedia learning with web lectures. Technical report, Georgia Institute of Technology, 2006.

[26] Joost CF De Winter. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, 18(10), 2013.

[27] Chris Dede, Marilyn C Salzman, R Bowen Loftin, and Debra Sprague. Multisensory immersion as a modeling environment for learning complex scientific concepts. In *Modeling and simulation in science and mathematics education*, pages 282–319. Springer, 1999.

[28] Chris Dede, Tina A Grotzer, Amy Kamarainen, and Shari Metcalf. Ecoxpt: Designing for deeper learning through experimentation in an immersive virtual ecosystem. *Journal of Educational Technology & Society*, 20(4):166–178, 2017.

[29] Chris Dede, Tina A Grotzer, Amy Kamarainen, and Shari J Metcalf. Virtual reality as an immersive medium for authentic simulations. In *Virtual, Augmented, and Mixed Realities in Education*, pages 133–156. Springer, 2017.

[30] M Luisa Dematte, Robert Österbauer, and Charles Spence. Olfactory cues modulate facial attractiveness. *Chemical Senses*, 32(6):603–610, 2007.

[31] Norman K Denzin and Yvonna S Lincoln. *The Sage handbook of qualitative research*. Sage, 2011.

[32] Huong Q Dinh, Neff Walker, Larry F Hodges, Chang Song, and Akira Kobayashi. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. In *Virtual Reality, 1999. Proceedings., IEEE*, pages 222–228. IEEE, 1999.

[33] Jacob Donley, Christian Ritz, and Muawiyath Shujau. Analysing the quality of experience of multisensory media from measurements of physiological responses. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 286–291. IEEE, 2014.

[34] Catherine Eberbach and Kevin Crowley. From everyday to scientific observation: How children learn to observe the biologist's world. *Review of Educational Research*, 79(1):39–68, 2009.

[35] Laura Ermi and Frans Mäyrä. Fundamental components of the gameplay experience: Analysing immersion. *Worlds in play: International perspectives on digital games research*, 37(2):37–53, 2005.

[36] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L Cox. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2985–2994. ACM, 2014.

[37] Eric Fassbender. Atmosphaeres: Immersive 360 stock videos and vr experiences. URL `https://www.atmosphaeres.com/`.

[38] Csikszentmihalyi Flow. The psychology of optimal experience. *Harper&Row, New York*, 1990.

[39] Bent Flyvbjerg. Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European planning studies*, 16(1):3–21, 2008.

[40] Floyd J Fowler. *Improving survey questions: Design and evaluation*, volume 38. Sage, 1995.

[41] Christopher Frayling. Research in art and design. 1993.

[42] Waka Fujisaki, Midori Tokita, and Kenji Kariya. Perception of the material properties of wood based on vision, audition, and touch. *Vision research*, 109: 185–200, 2015.

[43] William Gaver. What should we expect from research through design? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 937–946. ACM, 2012.

[44] Lynda Joy Gerry. Paint with me: Stimulating creativity and empathy while painting with a painter in virtual reality. *IEEE transactions on visualization and computer graphics*, 23(4):1418–1426, 2017.

[45] Gheorghita Ghinea, Christian Timmerer, Weisi Lin, and Stephen R Gulliver. Mulsemedia: State of the art, perspectives, and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11 (1s):17, 2014.

[46] Marie H Giard and Franck Peronnet. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience*, 11(5):473–490, 1999.

[47] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

[48] B Strauss Glaser and Anselm Strauss. The discovery of grounded theory : strategies for qualitative research. *Chicago : Aldine Publishing*, 1967.

[49] Scott W Greenwald, Wiley Corning, and Pattie Maes. Multi-user framework for collaboration and co-creation in virtual reality. 12th International Conference on Computer Supported Collaborative Learning (CSCL), 2017.

[50] Scott W Greenwald, Alexander Kulik, André Kunert, Stephan Beck, Bernd Fröhlich, Sue Cobb, Sarah Parsons, Nigel Newbutt, Christine Gouveia, Claire Cook, et al. Technology and applications for collaborative learning in virtual reality. Philadelphia, PA: International Society of the Learning Sciences., 2017.

[51] Jan Gugenheimer, Evgeny Stemasov, Julian Frommel, and Enrico Rukzio. Sharevr: Enabling co-located experiences for virtual reality between hmd and non-hmd users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4021–4033. ACM, 2017.

[52] Michael Haverkamp. Application of synesthetic design as multi-sensory approach on sound quality, 2001.

[53] Cindy E Hmelo-Silver, Surabhi Marathe, and Lei Liu. Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *The Journal of the Learning Sciences*, 16(3):307–331, 2007.

[54] David Hoptman. Invirovr davidhoptman's videos. URL `https://vimeo.com/invirovr/videos`.

[55] Hikaru Ibayashi, Yuta Sugiura, Daisuke Sakamoto, Natsuki Miyata, Mitsunori Tada, Takashi Okuma, Takeshi Kurata, Masaaki Mochimaru, and Takeo Igarashi. Dollhouse vr: a multi-view, multi-user collaborative design workspace with vr technology. In *SIGGRAPH Asia 2015 Emerging Technologies*, page 8. ACM, 2015.

[56] Ali Israr, Siyan Zhao, Kaitlyn Schwalje, Roberta Klatzky, and Jill Lehman. Feel effects: enriching storytelling with haptic feedback. *ACM Transactions on Applied Perception (TAP)*, 11(3):11, 2014.

[57] Hiroo Iwata, Hiroaki Yano, Takahiro Uemura, and Tetsuro Moriya. Food simulator: A haptic interface for biting. In *Virtual Reality, 2004. Proceedings. IEEE*, pages 51–57. IEEE, 2004.

[58] Corey Brian Jackson, Kevin Crowston, Gabriel Mugar, and Carsten Osterlund. "guess what! you're the first to see this event": Increasing contribution to online production communities. In *Proceedings of the 19th International Conference on Supporting Group Work*, GROUP '16, pages 171–179, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4276-6. doi: 10.1145/2957276.2957284. URL http://doi.acm.org/10.1145/2957276.2957284.

[59] Randolph L Jackson and Eileen Fagan. Collaboration and learning within immersive virtual reality. In *Proceedings of the third international conference on Collaborative virtual environments*, pages 83–92. ACM, 2000.

[60] Dhruv Jain, Misha Sra, Jingru Guo, Rodrigo Marques, Raymond Wu, Justin Chiu, and Chris Schmandt. Immersive terrestrial scuba diving using virtual reality. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1563–1569. ACM, 2016.

[61] Shunichi Kasahara and Jun Rekimoto. Jackin head: immersive visual telepresence system with omnidirectional wearable camera for remote collaboration. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, pages 217–225. ACM, 2015.

[62] Kathswinbourne. The smell of the australian outback, Jun 2013. URL http://kathswinbourne.com.au/the-smell-of-the-australian-outback/.

[63] Karlene Kerfoot. Learning intuition-less college and more kindergarten: The leader's challenge. *Dermatology Nursing*, 15(6):567, 2003.

[64] Prabal Khanal, Akshay Vankipuram, Aaron Ashby, Mithra Vankipuram, Ashish Gupta, Denise Drumm-Gurnee, Karen Josey, Linda Tinker, and Marshall Smith. Collaborative virtual reality based advanced cardiac life support training simulator using virtual reality principles. *Journal of biomedical informatics*, 51:49–59, 2014.

[65] Kristian Kiili, Timo Lainema, Sara de Freitas, and Sylvester Arnab. Flow framework for analyzing the quality of educational games. *Entertainment Computing*, 5(4):367–377, 2014.

[66] Konstantina Kilteni, Ilias Bergstrom, and Mel Slater. Drumming in immersive virtual reality: the body shapes the way we play. *IEEE transactions on visualization and computer graphics*, 19(4):597–605, 2013.

[67] June Kim and Tomasz Bednarz. Virtual reality to save endangered animals: Many eyes on the wild. In *2017 IEEE Virtual Reality (VR)*, pages 436–436. IEEE, 2017.

[68] Jane E Klobas, Stefano Renzi, and Maria Luisa Nigrelli. A scale for the measurement of self-efficacy for learning (sel) at university. 2007.

[69] Jonathan Knispel and Fraser Bullock. Collaborative vr painting in web browsers. In *SIGGRAPH Asia 2017 VR Showcase*, page 4. ACM, 2017.

[70] Günther Knoblich, Stephen Butterfill, and Natalie Sebanz. Psychological research on joint action: theory and data. In *Psychology of learning and motivation*, volume 54, pages 59–101. Elsevier, 2011.

[71] Duane V Knudson. *Qualitative diagnosis of human movement: improving performance in sport and exercise*. Human kinetics, 2013.

[72] Babis Koniaris, Ivan Huerta, Maggie Kosek, Karen Darragh, Charles Malleson, Joanna Jamrozy, Nick Swafford, Jose Guitian, Bochang Moon, Ali Israr, et al. Iridium: immersive rendered interactive deep media. In *ACM SIGGRAPH 2016 VR Village*, page 11. ACM, 2016.

[73] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[74] Aradhna Krishna, Ryan S Elder, and Cindy Caldara. Feminine to smell but masculine to touch? multisensory congruence and its effect on the aesthetic experience. *Journal of Consumer Psychology*, 20(4):410–418, 2010.

[75] Eric Krokos, Catherine Plaisant, and Amitabh Varshney. Virtual memory palaces: immersion aids recall. *Virtual Reality*, 23(1):1–15, 2019.

[76] Quang Tuan Le, Akeem Pedro, and Chan Sik Park. A social virtual reality based construction safety education system for experiential learning. *Journal of Intelligent & Robotic Systems*, 79(3-4):487–506, 2015.

[77] Johannes Leder, Tina Horlitz, Patrick Puschmann, Volker Wittstock, and Astrid Schütz. Comparing immersive virtual reality and powerpoint as methods for delivering safety training: Impacts on risk perception, learning, and decision making. *Safety Science*, 111:271–286, 2019.

[78] Ju-Hwan Lee and Charles Spence. Assessing the benefits of multimodal feedback on dual-task performance under demanding conditions. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1*, pages 185–192. British Computer Society, 2008.

[79] Pedro Lopes, Alexandra Ion, and Patrick Baudisch. Impacto: Simulating physical impact by combining tactile stimulation with electrical muscle stimulation. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 11–19. ACM, 2015.

[80] Jonas Löwgren. Annotated portfolios and other forms of intermediate-level knowledge. *Interactions*, 20(1):30–34, 2013.

[81] José Martins, Ramiro Gonçalves, Frederico Branco, Luís Barbosa, Miguel Melo, and Maximino Bessa. A multisensory virtual experience model for thematic tourism: A port wine tourism application proposal. *Journal of destination marketing & management*, 6(2):103–109, 2017.

[82] Terrence Masson, Ken Perlin, et al. Holo-doodle: an adaptation and expansion of collaborative holojam virtual reality. In *ACM SIGGRAPH 2017 VR Village*, page 9. ACM, 2017.

[83] Mathias. Water quality monitoring program. URL `http://www.ansbookshop.org/index.php/water-quality-monitoring-program`.

[84] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[85] Duncan C McKinley, Abe J Miller-Rushing, Heidi L Ballard, Rick Bonney, Hutch Brown, Susan C Cook-Patton, Daniel M Evans, Rebecca A French, Julia K Parrish, Tina B Phillips, et al. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, 208:15–28, 2017.

[86] James C McKinney. The diagnosis and correction of vocal faults: a manual for teachers of singing and for choir directors (revised ed.). *Nashville: Genevox*, 1994.

[87] A McMahan. Immersion, engagement, and presence: A method for analysing 3-d video games the video game theory reader'. *The video game theory reader, Routledge, New York, NY, US*, 2003.

[88] Mimi. Macaron troubleshooting guide, Mar 2018. URL `https://www.indulgewithmimi.com/macaron-troubleshooting-guide/`.

[89] Teresa Monahan, Gavin McArdle, and Michela Bertolotto. Virtual reality for collaborative e-learning. *Computers & Education*, 50(4):1339–1353, 2008.

[90] Janice M Morse. Using qualitative methods to access the pain experience. *British journal of pain*, 9(1):26–31, 2015.

[91] Tomoyuki Mukasa, Shintaro Murakami, and Tony Tung. Virtual interior stylization. In *ACM SIGGRAPH 2016 Appy Hour*, page 5. ACM, 2016.

[92] Randall Munroe. Xkcd. URL `https://xkcd.com`.

[93] Shintaro Murakami, Tomoyuki Mukasa, and Tony Tung. Mobile virtual interior stylization from scale estimation. In *ACM SIGGRAPH 2016 Posters*, page 14. ACM, 2016.

[94] Marianna Obrist, Carlos Velasco, Chi Vi, Nimesha Ranasinghe, Ali Israr, Adrian Cheok, Charles Spence, and Ponnampalam Gopalakrishnakone. Sensing the future of hci: touch, taste, and smell user interfaces. *interactions*, 23 (5):40–49, 2016.

[95] Biswaksen Patnaik, Andrea Batch, and Niklas Elmqvist. Information olfactation: Harnessing scent to convey data. *IEEE Transactions on Visualization Computer Graphics*, 2019. URL `http://www.umiacs.umd.edu/~elm/projects/info-olfac/info-olfac.pdf,PDFhttps://doi.org/10.1109/TVCG.2018.2865237,DOI`.

[96] Michael Q Patton. Qualitative research methods and evaluation. *Methods, Sage, Thousand Oaks*, 2002.

[97] Philippe Peigneux, Steven Laureys, Xavier Delbeuck, and Pierre Maquet. Sleeping brain, learning brain. the role of sleep for memory systems. *Neuroreport*, 12(18):A111–A124, 2001.

[98] William D Perreault and Laurence E Leigh. Reliability of nominal data based on qualitative judgments. *Journal of marketing research*, 26(2):135, 1989.

[99] Susan Persky and Jim Blascovich. Immersive virtual video game play and presence: Influences on aggressive feelings and behavior. *Presence: Teleoperators and Virtual Environments*, 17(1):57–72, 2008.

[100] Andrew M Pomerantz. *Clinical psychology: Science, practice, and culture*, chapter 7, pages 151–181. Sage Publications, 2016.

[101] Greg Pond. *Personal Communication*. Environmental Protection Agency, 2015-2017.

[102] Gregory J Pond, Margaret E Passmore, Frank A Borsuk, Lou Reynolds, and Carole J Rose. Downstream effects of mountaintop coal mining: comparing biological conditions using family-and genus-level macroinvertebrate bioassessment tools. *Journal of the North American Benthological Society*, 27(3):717–737, 2008.

[103] Jenny Preece, Helen Sharp, and Yvonne Rogers. Interaction design: beyond human-computer interaction. 2016.

[104] Joseph Psotka. Immersive training systems: Virtual reality and education and training. *Instructional science*, 23(5-6):405–431, 1995.

[105] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. Galaxy zoo: Exploring the motivations of citizen science volunteers. *arXiv preprint arXiv:0909.2925*, 2009.

[106] João Ramalho and Teresa Chambel. Immersive 360 mobile video with an emotional perspective. In *Proceedings of the 2013 ACM international workshop on Immersive media experiences*, pages 35–40. ACM, 2013.

[107] Nimesha Ranasinghe, Pravar Jain, Shienny Karwita, David Tolley, and Ellen Yi-Luen Do. Ambiotherm: enhancing sense of presence in virtual reality by simulating real-world environmental conditions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1731–1742. ACM, 2017.

[108] Armin Rest, DL Welch, NB Suntzeff, L Oaster, H Lanning, K Olsen, RC Smith, AC Becker, M Bergmann, P Challis, et al. Scattered-light echoes from the historical galactic supernovae cassiopeia a and tycho (sn 1572). *The Astrophysical Journal Letters*, 681(2):L81, 2008.

[109] Hauke Riesch, Clive Potter, and Linda Davies. Combining citizen science and public engagement: the open airlaboratories programme. *JCOM*, 12(3):A03, 2013.

[110] Harry V Roberts. Stock-market patterns and financial analysis: Methodological suggestions. *The Journal of Finance*, 14(1):1–10, 1959.

[111] Doug Rohrer and Harold Pashler. Increasing retention without increasing study time. *Current Directions in Psychological Science*, 16(4):183–186, 2007.

[112] Jeremy Roschelle. Learning by collaborating: Convergent conceptual change. *The journal of the learning sciences*, 2(3):235–276, 1992.

[113] Nancy E Roth and Wayne S Davis. *Biological indicator variability and stream monitoring program integration: a Maryland case study.* US Environmental Protection Agency, 2002.

[114] Ladan Shams and Aaron R Seitz. Benefits of multisensory learning. *Trends in cognitive sciences*, 12(11):411–417, 2008.

[115] Charles Spence, Marianna Obrist, Carlos Velasco, and Nimesha Ranasinghe. Digitizing the chemical senses: Possibilities & pitfalls. *International Journal of Human-Computer Studies*, 107:62–74, 2017.

[116] Robert Stickgold and Jeffrey M Ellenbogen. Quiet! sleeping brain at work. *Scientific American Mind*, 19(4):22–29, 2008.

[117] Erik Stolterman. The nature of design practice and implications for interaction design research. *International Journal of Design*, 2(1), 2008.

[118] Alina Striner and Anna Borovikov. Embodied vr experience to help non-scientists experience the subtle, gradual and cumulative nature of climate change, Mar 2017.

[119] Alina Striner and Jenny Preece. Streambed: Training citizen scientists to make qualitative judgments using embodied virtual reality training. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1201–1207. ACM, 2016.

[120] S Tretsiakova-McNally, E Maranne, F Verbecke, and V Molkov. Mixed e-learning and virtual reality pedagogical approach for innovative hydrogen safety training of first responders. *International Journal of Hydrogen Energy*, 42(11):7504–7512, 2017.

[121] Cordula Vesper, Ekaterina Abramova, Judith Bütepage, Francesca Ciardo, Benjamin Crossey, Alfred Effenberg, Dayana Hristova, April Karlinsky, Luke McEllin, Sari RR Nijssen, et al. Joint action: mental representations, shared information and general mechanisms for coordinating with others. *Frontiers in Psychology*, 7:2039, 2017.

[122] Ian Warren. Teaching patterns and software design. In *Proceedings of the 7th Australasian conference on Computing education-Volume 42*, pages 39–49. Australian Computer Society, Inc., 2005.

[123] Andrea Wiggins. Free as in puppies: compensating for ict constraints in citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1469–1480. ACM, 2013.

[124] Andrea Wiggins, Greg Newman, Robert D. Stevenson, and Kevin Crowston. Mechanisms for data quality and validation in citizen science. In *Proceedings of the 2011 IEEE Seventh International Conference on e-Science Workshops*, ESCIENCEW '11, pages 14–19, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4598-1. doi: 10.1109/eScienceW.2011.27. URL http://dx.doi.org/10.1109/eScienceW.2011.27.

[125] William Winn, Mark Windschitl, Ruth Fruland, and Yenling Lee. When does immersion in a virtual environment help students construct understanding.

[126] Cathy Wiss. *How to Read Your Stream*. Maryland Audubon Society, 2014. URL http://md.audubon.org/.

[127] Christiane Wölfel and Timothy Merritt. Method card design dimensions: a survey of card-based design tools. In *IFIP Conference on Human-Computer Interaction*, pages 479–486. Springer, 2013.

[128] Nesra Yannier, Scott E Hudson, Eliane Stampfer Wiese, and Kenneth R Koedinger. Adding physical objects to an interactive game improves learning and enjoyment: Evidence from earthshake. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):26, 2016.

[129] Masasuke Yasumoto and Takehiro Teraoka. Shadow shooter: 360-degree all-around virtual 3d interactive content. In *ACM SIGGRAPH 2015 Posters*, page 29. ACM, 2015.

[130] Scott Young. *Holistic Learning (e-book)*, 2017 (accessed February 2, 2017). URL `http://www.scotthyoung.com/blog/Programs/HolisticLearningEBook.pdf`.

[131] John Zimmerman and Jodi Forlizzi. The role of design artifacts in design theory construction. *Artifact*, 2(1):41–45, 2008.

[132] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502. ACM, 2007.

[133] Longhao Zou, Irina Tal, Alexandra Covaci, Eva Ibarrola, Gheorghita Ghinea, and Gabriel-Miro Muntean. Can multisensorial media improve learner experience? In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 315–320. ACM, 2017.