Active Prolonged Engagement EXpanded (APEX): A Toolkit for Supporting Evidence-Based Iterative Design Decisions for Collaborative, Embodied Museum Exhibits

DURI LONG, Expressive Machinery Lab, Georgia Institute of Technology, Atlanta, USA
TOM MCKLIN, The Findings Group LLC, Decatur, USA
NYLAH AKUA ADJEI BOONE, Georgia Institute of Technology, Atlanta, USA
DEZARAE DEAN, Georgia Institute of Technology, Atlanta, USA
MIRINA GAROUFALIDIS, Georgia Institute of Technology, Atlanta, USA
BRIAN MAGERKO, Expressive Machinery Lab, Georgia Institute of Technology, Atlanta, USA

This article presents Active Prolonged Engagement eXpanded (APEX), a framework and toolkit for informing evidence-based decisions about the iterative design of embodied, collaborative museum exhibits. We provide an overview of APEX, a framework that builds on both prior work and experimentally derived data to provide an understanding of how visitors' physical, social, emotional, and intellectual engagement transform during the course of their interaction with an exhibit. We present two case studies demonstrating how to apply APEX in practice, analyzing video recordings of participant interactions with different design iterations of *TuneTable*—an interactive exhibit for co-creative computational music-making—at both a macro- and micro-level. In the case studies, we explore how APEX reveals important features of participant interaction that suggest implications and directions for design. Finally, we present a toolkit of resources to aid researchers in operationalizing APEX as a framework for video analysis, in-situ observation, and iterative design and evaluation.[1]

CCS Concepts: • **Social and professional topics → Professional topics → Computing education → Informal education • Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods; • Human-centered computing → Collaborative and social computing → Collaborative and social computing design and evaluation methods; • Applied computing → Education → Collaborative learning**

Additional Key Words and Phrases: Museum exhibits, informal learning, co-creativity, embodiment, tangible, evaluation, qualitative analysis, collaborative learning, family group learning, CS education

ACM Reference format:

Duri Long, Tom McKlin, Nylah Akua Adjei Boone, Dezarae Dean, Mirina Groufalidis, and Brian Magerko. 2022. Active Prolonged Engagement Expanded (APEX): A Toolkit for Supporting Evidence-Based Iterative Design Decisions for Collaborative, Embodied Museum Exhibits. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 6, CSCW1, Article 50 (April 2022), 33 pages, https://doi.org/10.1145/3512897

This work is supported by the National Science Foundation, under grant DRL-1612644.

2573-0142/2022/04 – Article#50… $15.00

https://doi.org/10.1145/3512897

1 INTRODUCTION

Creating educational experiences for informal learning spaces—such as museums, science centers, or after-school centers—introduces a number of design constraints and opportunities that do not exist in formal learning environments. Informal spaces are often "free-choice" learning environments in which visitors structure their own learning experiences and can choose how long and in what depth they want to engage in different activities [22]. There are many exhibits or activities for groups to choose from, and visitors may be easily distracted or have little patience for activities that are difficult to grasp [2]. People visit these spaces in groups (e.g. family groups, school trips) as a leisure activity, meaning that activities excel when they are engaging, fun, and maximize opportunities for collaboration [30]. "Learning" in these environments is not strictly focused on content-knowledge gain, but also deals with socio-emotional factors such as improving perceptions of a field or topic, fostering interest development, and providing memorable and awe-inspiring experiences [7,22]. Learning experiences also often involve embodied interaction (e.g. [36,51]), which contrasts with the passive transmission model of learning that is found in many classroom environments.

Due to the unique features of informal learning, it is often difficult to assess learning interventions in museum spaces in the same manner we are accustomed to evaluating in a classroom. For example, pre/post tests are not well-suited to assessing learning for a visitor group that spends five minutes at an exhibit in a science center—both due to the short period of time between the "pre" and the "post" and because visitors having a fun afternoon at the museum may simply not want to take a test [2]. In addition, such metrics may only assess content-knowledge gain, ignoring important socio-emotional factors.

Despite the challenges of evaluating learning in informal spaces, it is important to do so. Evidence-based evaluations can contribute to a better understanding of which exhibits are the most engaging, which are contributing to interest development and learning, and which should be preserved or abandoned during transformative museum planning [3]. In addition, evaluation is crucial during the initial period of exhibit design and development. Assessing prototypes on the museum floor can lead to unexpected results and it is important to understand which design conditions contribute to the strengths and weaknesses of an exhibit [18,38,64]. Evidence-based assessments can aid in facilitating constructive iterative design of exhibits.

Researchers have developed methods for evaluating learning in museums, including frameworks for video analysis and live observation of participant interactions [6,39], retrospective interviews or post-tests specifically designed for the museum environment [23], and protocols for analyzing participant dialogue for evidence of learning [3,38,61]. However, there is still much room to explore in this space, particularly in regards to developing evidence-based tools to understand the unique relationship between content-knowledge gain, socio-emotional factors such as interest development and collaboration, and embodied interaction with exhibit components.

The central research question we explore in this paper is: How can we inform evidence-based decisions about the iterative design of interactive technology exhibits meant for collaborative, embodied active prolonged engagement [39] in informal learning spaces? We explore this question by reflecting on our own experience studying visitor interactions with TuneTable, a museum exhibit that aims to foster interest in computer science and communicate key computing concepts by engaging students in co-creative music programming. We begin by discussing related research on understanding and evaluating learning and interest development in museum spaces. We then describe the TuneTable exhibit. We then present APEX, a novel

framework for video analysis and observation in informal learning spaces that was derived from two previously existing frameworks [6,39] and a deductive thematic analysis of visitor interactions with Blockhead, an iteration of TuneTable. We apply APEX to two iterations of TuneTable, presenting each as a case study. The design goal underlying both iterations was fostering interest development and learning in computer science (CS) through co-creative music programming. However, the two iterations address the design goal in very different ways and thus represent two distinctly different exhibit designs. In the case studies, we explore how APEX reveals important features of participant interaction that suggest new implications and directions for design. We have recently expanded the APEX resources to include an observation form for lower-resource in-situ observation and a worksheet to aid designers in setting evaluation targets using APEX. We present these resources and then conclude by discussing limitations of the framework and directions for future work.

The core contributions of this paper include: 1) APEX, a framework that can be used to qualitatively understand participants' social, emotional, physical, and intellectual engagement with certain museum exhibits; 2) case studies demonstrating how to apply the APEX framework to understand participant interactions with two different museum exhibits; and 3) a toolkit of supplemental resources to aid researchers in operationalizing APEX in their design process, data collection, and analysis. These resources can aid researchers in the CSCW community in designing and evaluating collaborative, embodied museum exhibits with interactive technology.

2 RELATED WORK

2.1 Learning in Museums

Theories of *constructivism* are important to our understanding of museums and other informal learning spaces. Constructivism argues that learners actively engage with their environments to construct knowledge [7,32]. Among the theoretical origins of constructivism are Piaget's work on the experiences of learners, as well as the conclusions of Vygotsky and Bruner, which frame all knowledge as socially constructed [4]. For these theorists and others, knowledge is constructed actively and socially among learners in an environment [4,14,42,58,76]. The design of contemporary museums in particular is especially indebted to Dewey, who championed the application of social constructivism to public education [32].

Viewed through the lens of social constructivism, *learning* is an active process of meaning-making that involves connections to both social and personal identity. We define learning using social constructivism because of its impact on the design of museum and science centers—especially on those that operate as *free-choice* learning environments, wherein learner agency is emphasized and exhibits facilitate embodied social interaction [22]. This stands in opposition to the "object-based epistemology" which guided museum design in the nineteenth century and which was focused on presenting objects/artifacts as knowledge, with little or no focus on the visitor and the prior knowledge/experiences that they brought to the museum [7,17].

2.2 CSCW and Museum Learning

The social, cooperative nature of learning in museums combined with the growing importance of technology-driven interactives in museum spaces has led to an interest in museum exhibit design and methods for understanding visitor interaction in the CSCW community. Researchers in the CSCW community have investigated how to design collaborative museum exhibits (e.g.

[36,50,61,63]), develop design principles to better support group learning [24,74], better understand social behaviors at museum exhibits [40,50], and engage in co-designing exhibits with visitors [16,25,44]. Research related to studying social interaction with large public displays and art interactives (e.g.[8,10,13,49,57,79]) has also informed our work. In particular, the *physical* and *social* dimensions of APEX draw on research that explores the different types of social roles that people take on and people's negotiation of physical space surrounding public interactives.

Other relevant research has investigated how to develop frameworks for evaluating collaborative usability in workspaces. For example, Pinelle and Gutwin present a technique for evaluating collaborative usability on tabletop workspace interfaces [59]. This framework shares some characteristics with APEX—such as the focus on identifying interactions spanning physical, social, and verbal modes of interaction. However, APEX differs from the framework presented by Pinelle and Gutwin in several significant ways: 1) APEX is designed for learning environments (not workspaces) and therefore focuses on representing visitors' progression through different stages of interaction, rather than on identifying overall usability issues; 2) APEX can be used to explore different patterns of participant engagement in a non-prescriptive way and allows design teams to choose what dimensions of engagement to focus on and decide what an "ideal" interaction looks like; and 3) although we developed APEX through analysis of tabletop interfaces, we have tested it with other exhibits and it is intended to be more broadly useful for collaborative, embodied museum exhibits.

Most relevant to the work in this paper are researchers in CSCW and closely related communities (e.g. CSCL, CHI) who have developed methods for better understanding visitor interaction and learning at collaborative, embodied museum exhibits. In particular, we discuss and draw on Roberts and Lyons' framework for analyzing social learning talk at exhibits [61] and Hornecker and Ciolfi's extensive research and experience on designing and evaluating collaborative, embodied, technology-centered museum exhibits [38].

2.3 Evaluation in Museums

There is still some ambiguity regarding how to best evaluate free-choice learning experiences. It is clear that the goals of museums extend beyond encouraging content-knowledge gain—in the spirit of constructivism, they also aim to provide experiences that promote social interaction and identity construction [21,34,65] and emotional connections with the learning material [26]. Researchers in visitor studies have developed a variety of approaches to address the issue of assessing learning in informal spaces, and several sources provide a detailed overview of evaluation methods used in museums (c.f. [3,38]). We summarize this work here, and organize this section using the framework outlined in a 2008 NSF report in which prominent researchers in visitor studies identified five key areas to focus on when evaluating exhibits: *knowledge, engagement, attitude, behavior,* and *skills* [3].

Knowledge deals with visitors' understanding of content knowledge [3]. Pre/post-tests can serve as effective tools for assessing knowledge, but they are often too lengthy for practical deployment in a museum setting where visitors have limited time. Testing participants' ability to articulate the key idea of an exhibit can serve as a speedier, more informal "post-test" [22]. Others have asked visitors to draw concept maps before and after engaging in an activity as a more open-ended method of assessing knowledge, allowing for a variety of learning outcomes depending on the visitor group and their motivations [23]. Observational methods such as

listening for visitor discussion of relevant content [1,61] are also used as knowledge assessments.

Engagement is defined as visitors' "excitement and involvement" with the exhibit and associated content knowledge [3]. Engagement is often measured using hold time (i.e. the amount of time participants spend at an exhibit). Conversational indicators such as curiosity-related language [60] can help researchers to understand how visitors engaged with an exhibit. Asking participants why they left or disengaged with an exhibit can also provide insight into barriers to engagement [9]. For technology-based installations, interaction logs that track what content visitors interact with and for how long can provide an easily scalable metric of engagement with content knowledge [38,41]. Some have studied engagement with multiple exhibits by tracking visitors' movements throughout the museum [43,69].

Attitude overlaps somewhat with engagement, but additionally deals with longer-term perceptions of the exhibit and its associated subject matter [3]. Positive attitudes towards a subject can play an important role in interest development [33]. Participants can be asked to self-report their attitude in a post-interaction survey [3], and observable markers such as emotional reactions [72] can also be used as indicators of potential attitudinal change.

Behavior describes how museum exhibits influence visitors' lives after they leave the museum. Longitudinal studies of visitors' experiences before, during, and after their museum visits can provide a detailed picture of the effects informal learning has on behavior (e.g. [19]). However, longitudinal studies can be infeasible in many cases. As a result, many museum researchers ask visitors to self-report their intent-to-persist in the subject matter as they are leaving the museum or via a survey/interview shortly after the visit as a way of assessing behavior [3].

Finally, *skills* deals with "the procedural aspects of knowing" [3]. Conversational analysis can be used to illuminate when participants are developing inquiry skills and engaging in "learning-talk" [1,61]. Physical skill development is also important, but there is an identified need for additional work on evaluating physical skills [2]. Some researchers have used tools like Goodwin's Embodied Participation Framework to analyze embodied interaction and communication at exhibits at a micro-level, looking at factors such as posture, body alignment, head movement, gesture, gaze, and spoken language [31,68]. Researchers interested in understanding tangible interaction have also used techniques such as interaction logging and unobtrusive observation to understand physical interactions [35,38].

These five areas of evaluation (*knowledge*, *engagement*, *attitude*, *behavior*, *skills*) make it clear that a mixed-methods approach is necessary in order to fully understand informal learning experiences. Some of the methods used to measure each area (e.g. longitudinal studies, post-tests) provide rich and valuable data but cannot feasibly be conducted with large numbers of participants. The methodologies that are most useful in practice often rely instead on unobtrusive observation techniques. Our work explores how to develop a framework for unobtrusive observation and video analysis that can provide an integrative view of visitors' physical, social, emotional, and intellectual engagement with an exhibit over time. Our framework is intended to contribute to the ecosystem of existing evaluation methods, and can be used in concert with other techniques depending on researchers' interests/needs.

3 THE EXHIBITS

We originally developed the APEX framework through our work on *TuneTable,* a project in which we aimed to design a tangible tabletop museum exhibit for co-creative music programming. Co-creative computing activities have previously been shown to foster interest development and self-efficacy in computing, particularly for historically marginalized groups [15,27,52,77]. *TuneTable* aimed to leverage these findings to develop a museum exhibit that facilitated an engaging, co-creative interaction with computing in order to foster interest development.

The APEX framework initially emerged from a deductive thematic analysis of participant interactions with an early iteration of *TuneTable* called *Blockhead,* guided by two existing frameworks [6,39]. We later adapted and applied the framework to a second, very different iteration called *GrooveMachine*. This section describes the two exhibit iterations and the setups for our studies of both exhibits. Both iterations are designed to encourage collaboration and are targeted at family groups with middle school age children. *TuneTable* is in conversation with a variety of prior work that has explored how to design tangible interactives to foster collaborative learning (e.g. [35,37,54,66]—we do not review this work in detail here as the primary contribution of this paper is the APEX framework, not the exhibit designs, which have been described in previous publications [11,46,47]). The two different design directions represented by *Blockhead* and *GrooveMachine* were an effort on the part of the design team to explore a variety of approaches to embodied computing education experiences in a museum.

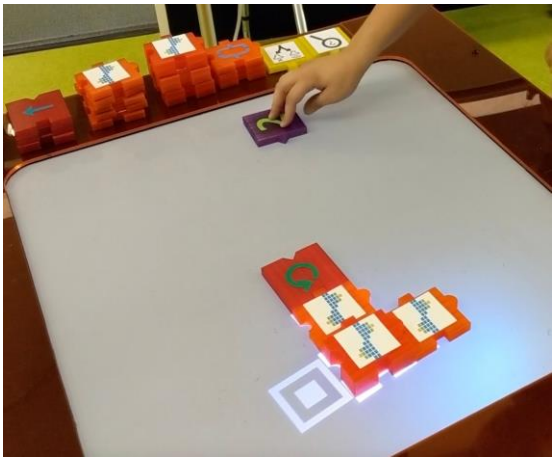


Figure 1. *Blockhead* iteration of the *TuneTable* project. In this image, several sound samples are chained together to make a tune. A loop function block is attached to one row of sound samples and causes them to repeat.

3.1 Blockhead

3.1.1 Description. Blockhead (Figure 1) is an interactive tabletop museum exhibit in which participants can co-create music using computing concepts. Visitors interact with *Blockhead* using a bespoke programming language in which puzzle-piece shaped tangible blocks are placed on the table. When a sample block is placed on the table, a 'play head' is spawned. Visitors can tap the play head with their finger to play the sound sample that the block is associated with. Sample blocks can be connected together to create a tune (i.e. a chain of consecutively played

music samples). Function blocks can also be added to chains to create a subroutine. Function blocks reflect common computing concepts such as loops, conditionals, and "go-to" statements. For example, a loop function block connected to a sample block would make the sound sample repeat. More detail on the functionality of the *Blockhead* programming language can be found in [47].

3.1.2 Study Setup. We observed participant interactions with *Blockhead* at the Museum of Science and Industry Chicago during two different data collection sessions in 2017. Members of our research team recruited a total of 31 groups of middle-school (i.e. 10-14 year old) children and their parents (112 participants in total) to interact with *Blockhead* by approaching family groups with children who appeared to be in the target age range as they entered the museum. Visitors were asked to interact with *Blockhead*, which was installed in a classroom workspace in the museum, and complete a short interview and survey after their interaction. Participant interactions with the exhibit were video recorded from a top and a side view. The study setup and participant demographics are discussed in more detail in [47].

3.2 GrooveMachine

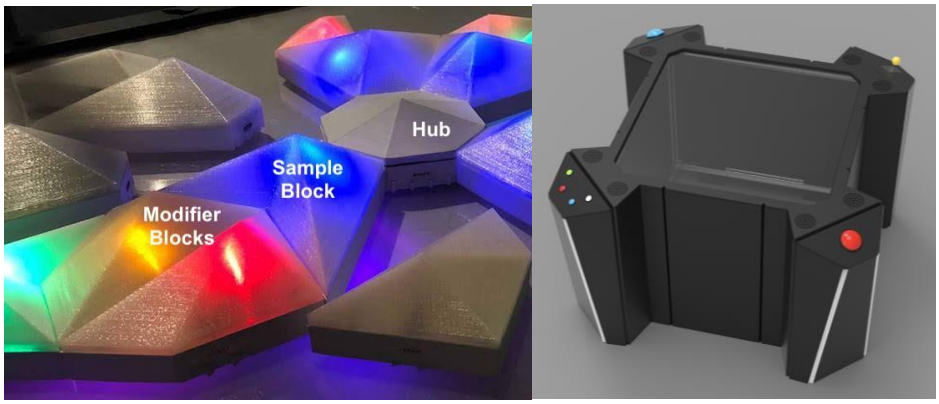


Figure 2. *GrooveMachine* iteration of *TuneTable* project. Left image is a close-up of the blocks on the tabletop. Right image is a mock-up of what the table structure looks like, with arcade buttons on each corner.

3.2.1 Description. GrooveMachine (Figure 2) diverges from the block-based coding syntax used in *Blockhead* and many other introductory computing environments and instead uses embodied metaphors to communicate computing concepts through music making / play. Users learn about fundamental computing concepts like loops, parameters, variable scope, and objects by engaging in tangible interactions with the exhibit that produce musical changes.

The table is divided into quadrants by its protruding corners (Figure 2). These four interaction stations were intended to provide individual workspaces for learners to enable them to work on their section of the group composition without being disrupted or interrupted by others. This was partially inspired by the *Spinning Blackboards* exhibit discussed in [39] and Hornecker et al.'s principle of multiple access points [37].

Visitors begin interacting with the table by adding sample blocks to a central hub (Figure 2). Blocks light up when they are connected to the hub. The system scans the hub in a continuous loop, playing either the sound associated with the sample block or a silent "rest" at each of the

eight "steps" around the hub. Lights in the central hub indicate where the step counter is at a given moment. This process is an embodied metaphor for a computational loop and the process of placing data in a loop. The blocks are also a metaphor for object oriented programming, as each sample piece is physically identical but contains different audio data.

Visitors can add up to four modifier blocks (Figure 2) to each sample block. Modifiers act like parameters, as they change the way data is processed without changing the data itself. Each modifier block has a different effect on the sound sample it is attached to—high orchestration, low orchestration, repetition or reversal.

Finally, each corner of the table has a set of arcade controls on it (Figure 2). The arcade controls affect the music on the table at a global scope, affecting factors such as volume, tempo, or sound distortion. These controls, combined with the modifiers, are an embodied metaphor for variable scope. The arcade controls are intended also as an entry point of engagement for younger kids. The design of *GrooveMachine* is described in more detail in [11].

3.2.2 Study Setup. We installed *GrooveMachine* "in-the-wild" on the museum floor for two days at the Museum of Science and Industry, Chicago during the summer of 2019. We used an implied consent procedure, placing several large signs surrounding the exhibit informing participants that they would be video recorded if they entered the space. Interacting with the exhibit constituted consent to the video recording. Video was recorded from both a top and a side view. After groups were finished with their interaction, groups appearing to have children in the target age range (10-14) were taken aside and asked for verbal consent to complete child interviews and parent surveys. We analyzed video data of 35 total visitor groups.

4 FOUNDATIONAL EVALUATION FRAMEWORKS

The research agenda from which this work emerged was not initially focused on developing new methods for evaluating museum exhibits. We were instead focused on designing a tangible tabletop exhibit for co-creative music programming. However, we quickly realized the need for a robust, empirically rigorous tool for assessing factors such as participant engagement, learning, and socio-emotional factors in order to inform our design process. We wanted to use a framework that allowed us to gain a holistic picture of participant engagement (i.e. socio-emotional factors in addition to content knowledge), could be reliably and consistently applied across multiple visitor groups and design iterations, and could be operationalized within a relatively short time frame (i.e. not a longitudinal study).

Conversational analysis frameworks such as Roberts and Lyons' learning talk framework [61] provided insight into visitors' content-related dialogue and group management dialogue at an embodied exhibit, but we were also interested in additional factors such as participants' physical interactions with the exhibit and indicators of emotional engagement. Interaction logging [38,41] is another technique we have used for evaluating participant engagement in our work, but as Hornecker points out, log files "cannot provide the reasons why visitors behaved the way they did, nor rich data documenting their reactions and thoughts" [38].

Our museum partner (the Museum of Science and Industry Chicago) suggested we look at two frameworks for unobtrusive observation that are commonly used to provide a holistic understanding of participant engagement in museum spaces: 1) Barriault and Pearson's Visitor Engagement Framework (VEF) [6] and 2) Humphrey et al.'s Active Prolonged Engagement (APE) framework [39]. Both APE and VEF focus on understanding learning and engagement at the group level, since museum exhibits are typically frequented by family or school groups. In

this section, we discuss each of the frameworks and some of the challenges we ran into when applying them to our analysis. We then discuss how these efforts informed the APEX framework we developed.

4.1 The Visitor Engagement Framework (VEF)

Barriault and Pearson's Visitor Engagement Framework (VEF) groups observable learning-related behaviors (e.g. explaining an exhibit to a friend, reading signage) into three stages—*initiation, transition,* and *breakthrough* behaviors [6]. *Initiation* includes behaviors such as doing the activity once or incompletely, or watching the activity; *transition* includes behaviors such as repeating the activity or expressing a positive emotional response; *breakthrough* includes behaviors such as referencing prior experiences, seeking and sharing information, or demonstrating inquiry behavior via experimentation [6]. Barriault and Pearson claim that the more visitors that reach the transition and breakthrough stages, the more the exhibit facilitates learning.

VEF has many strengths as a framework. It correlates observable behaviors with stages of progress in the learning process, making it a useful tool for assessing to what degree an exhibit facilitates learning. In addition, VEF can be used as a tool for quickly evaluating new exhibits or comparing existing exhibits with each other.

However, VEF is less useful for more in-depth understanding of embodied co-creative experiences in a research context. VEF only provides an understanding of *how many* participants reach a certain stage of engagement. This leaves out *when, in what order*, and *how* participants reach certain stages of engagement. An analysis conducted using VEF does not provide insight into how participants navigate varying stages of engagement over time, or what types of behaviors precede transitions between stages of engagement. A VEF analysis also does not illuminate the relationship between different types of engagement (e.g. social vs. physical).

4.2 Active Prolonged Engagement (APE)

VEF was designed to be a quick and easy way to assess learning at museum exhibits during live observation sessions. In contrast, Humphrey et al. and the Exploratorium's *Active Prolonged Engagement* (APE) [39] framework was designed to provide museum researchers with more detailed information that could feed back into the exhibit design process.

Many exhibits at the Exploratorium were originally developed as *planned discovery (PD)* exhibits, in which visitors were presented with a key content-knowledge related question that was answered via a short interactive experience. Exhibit designers at the Exploratorium later developed an interest in designing and studying "APE exhibits," where *(A)ctive* means that interaction with the exhibit was led by visitors; *(P)rolonged* means that visitors spent more time at these exhibits (relative to other exhibits—Humphrey et al. found that visitors tended to engage with APE exhibits three times as long as they did with PD exhibits [39]); and *(E)ngaged* means that visitors built on previous actions as they interacted with the installation [39]. APE exhibits were targeted less on specific content-knowledge gain than PD exhibits, and were instead more focused on fostering visitors' curiosity, imagination, and meaningful interactions [39]. The APE framework consists of descriptions of behavioral markers of four different types of engagement—*intellectual, social, physical,* and *emotional.*

One of the strengths of APE is that it emphasizes the roles that different components of engagement play in the learning process. For example, emotional engagement may not reflect content-knowledge understanding as reliably as intellectual engagement, but it does play an

important role in interest development. However, engagement is evaluated for the exhibit as a whole (e.g. "Most visitors engaged in meaningful intellectual engagement with this exhibit"), and this high-level view omits details about what behaviors lead to learning or how different types of engagement relate to each other during the course of the interaction.

APE's focus on visitor curiosity, imagination, and meaningful interactions was a great fit for our project, which was primarily geared towards fostering creativity and interest development in relation to computing. However, the process of adapting the APE framework to a new installation was not straightforward. We found that the APE codes as-written were too subjective to apply consistently to our analysis (e.g. "Level 3: basic meaningful engagement. Level 3 is what we look at and say to ourselves 'They've got it. This is acceptable. It is adequate" [20]). Examples given in the APE codebook were very exhibit-specific [20]. Transferring the codes to our exhibit was complicated by the fact that the APE coding schemes do not adhere strictly to observable actions and instead stray into coding elements that cannot be observed. For example, the Exploratorium's coding scheme for an exhibit titled *Spinning Patterns* uses "intentionality" as one of its codes for intellectual engagement [20], but we have found in practice that participants rarely express their intent clearly and often analysts are forced to guess whether or not a participant had a specific intent when executing an action.

5 ACTIVE PROLONGED ENGAGEMENT EXPANDED (APEX)

Our experience applying VEF and APE to understanding our exhibits revealed strengths of the frameworks, but also highlighted some areas in which they were lacking for our purposes. The remainder of this paper describes how we developed APEX (**A**ctive **P**rolonged **E**ngagement e**X**panded), a framework that builds on the strengths of VEF and APE, but addresses some of their shortcomings. APEX is a more readily transferable, concrete coding scheme that can provide insight into visitor interactions and stages of engagement at both a micro (moment-by-moment within an individual group) and macro (across multiple groups) level. The APEX coding scheme we developed as well as the video coding procedure are described in detail in this section.

5.1 What types of exhibits/visitors is APEX intended for?

We contend that the APEX framework can be useful for understanding visitor interactions with collaborative, embodied museum exhibits intended to foster active prolonged engagement. This means visitors should be interacting with the exhibits together as a group using a tangible or full-body interface, engaging with them over a (relatively) long period of time, and that visitor experiences can be shaped by their motivations and interests, meaning that outcomes often look different from group to group. APEX is **not** as well suited for understanding participant interactions with planned discovery exhibits or exhibits that involve primarily individual interaction, do not engage visitors in an embodied way, and/or are not intended to foster group dialogue.

Humphrey et al. apply their APE analysis to studying family and friend groups instead of individuals and school groups [39]. They choose not to focus on school groups because the interactions may be driven by teacher assignments. In addition, visitor studies research shows that people visit museums in family and friend groups and often learn as a unit [30]. We similarly focus on analysis of families at the group level, though APEX could be used to understand more open-ended school group interactions as well.

5.2 Methodology: Developing the Framework

After our attempts to apply existing coding schemes (APE, VEF) to our video analysis were unsuccessful, we took a different approach and began a deductive thematic analysis [12] of the *Blockhead* video data—that is, we iteratively developed codes and themes guided by APE and VEF. More specifically, our deductive analysis was shaped by the four high-level *types* of engagement identified by Humphrey et al. (*social*, *physical*, *intellectual*, *emotional*) as well as the notion of relating observable behaviors to *stages* of engagement (*initiation*, *transition*, *breakthrough*), as was done in VEF. Some of the more salient codes in the APE and VEF codebooks also factored into our deductive analysis. Nowell et al. provide a multi-step approach for conducting and reporting on thematic analysis, which we use to structure this section and guide our description of our thematic analysis [56].

5.2.1 Familiarizing Ourselves with the Data. We gained initial familiarity with the data by conducting participant studies and watching through videos as we attempted to apply APE and the VEF to *Blockhead*. Part of familiarizing ourselves with the data involved determining a unit of analysis. We used a one-zero sampling approach [71] to code the video data. Using this approach, interactions are broken down into time segments (in our case, we divided each video recording into ten second segments) and each code is given a '1' if it occurred during that time segment and a '0' if it did not occur. One-zero sampling has been shown to be both a reliable and valid method of behavior observation, correlating significantly with measures of actual frequency and duration while avoiding the associated difficulties of defining behavior initiation and termination [71]. We used a similar approach for coding *intellectual engagement* (discussed further below), except we used lines of verbal utterances as our unit of analysis rather than a fixed ten second time interval in order to avoid splitting thoughts across segments. Transcripts and line-by-line breakdowns were made by one analyst and then verified by a second analyst. Any discrepancies were resolved prior to coding. For all categories of engagement, multiple codes can be applied to a single segment (e.g. two types of physical engagement might occur during the same segment).

5.2.2 Generating Initial Codes. We began to code the videos for what we termed *atomic actions* within each one of the four APE categories of engagement—*intellectual*, *social*, *physical*, and *emotional*. We defined atomic actions as actions taken by a single person that could not be broken down into a series of sub-actions. Examples include moving a block, making an observation, or laughing.

5.2.3 Searching for Themes. We sorted our list of atomic actions into *composite actions* (or *themes*) that described higher-level action sequences that involved multiple group members (e.g. teaching a family member, collaborating to compose a tune together). Finally, we synthesized composite actions into overarching themes that described engagement at the group level (since most exhibits are frequented by family or school groups learning together).

Table 1. Example of how we turned one of the original *Blockhead* codes (P_Isolated_Manipulation) into a code that could be transferred to a new exhibits, like *GrooveMachine*.

Code: Isolated Manipulation; Category: Physical	
Original Blockhead Code	**Transferrable Code**
Participants are physically interacting with the exhibit, but only with isolated components. They are trying out components of the system and have not yet begun to execute sequences of related actions involving multiple exhibit components. This involves placing a new sound block on table, moving it around, and (possibly) pressing play to test this block.	*General:* Participants are physically interacting with the exhibit, initially testing/touching/trying out components in isolation to determine how individual pieces work. *Keywords:* trying out individual components, potentially unsure, disconnected, haphazard, isolated *Specifics (GrooveMachine):* turning a block, randomly using the joysticks and buttons in the corners, struggling to connect a sample block, struggling to connect a modifier block, attempting to get connected blocks to light up *Specifics (Blockhead):* placing a new sound block on table, moving a sound block around, pressing play or pause to test or "audition" a sound block. *Specifics (Sound Happening):* doing a single action as a way of initially testing the system (e.g. kicking a ball through the space or bouncing the ball once); playing with the balls in such a way that interactions are indistinguishable from just playing with balls without sound—it is not clear that the individual(s) are noticing the sound. *Specifics (Dive Trainer):* sitting down in chair for first time, holding wheel, hitting buttons in a random or non-methodical way

5.2.4 Reviewing Themes. We utilized inter-rater reliability [28,55] as a tool for refining and reviewing our themes. The ability for multiple analysts to apply codes reliably was particularly important to us because we wanted to ensure that our themes were based on observable behaviors, thus rectifying one of the issues we identified with the original APE framework. In addition, we wanted to provide a transferable framework that could be used evaluate other co-creative embodied installations in the future. This desire for future transferability increases the importance of developing themes that can be reliably coded for by multiple analysts. We began an iterative process of having multiple analysts code videos using the set of themes we had developed. At team meetings, we discussed a) whether we had achieved an acceptable level of inter-rater reliability; b) what discrepancies were inhibiting reliable coding; and c) whether the themes accurately represented the data and told a cohesive and informative story.

5.2.5 Transferring Coding Scheme to New Exhibits. After we developed the coding scheme for *Blockhead*, we wanted to ensure that it could be transferred to other exhibits and was not just useful for our particular context. This was particularly important to us as the original APE coding scheme was not readily transferable to new exhibits. We transferred the coding scheme we developed for *Blockhead* to analyzing visitor interactions with *GrooveMachine*. Although *GrooveMachine* was developed with a focus on similar learning outcomes, its form and design rationale were significantly different from *Blockhead* and thus it constituted a distinct exhibit.

We began the process of transferring the original APEX codes to *GrooveMachine* by breaking them down into three sections: *general definition*, *specifics*, and *keywords*. In the *general definition*, we provide a brief overarching description of the code that could be applied to any exhibit. The *specifics* section is customizable depending on the exhibit in question. We also include a set of *keywords* for each code as well as a set of notes and guiding questions (where relevant) to aid researchers in filling out examples specific to their exhibit for each code. We provide an example of how we broke down one of the codes in Table 1.

We recommend that researchers looking to adapt the coding scheme to their exhibit observe participants engaging with their exhibit after reviewing the coding scheme and use their observations to fill in the "specifics" section with examples. Specifics may need to be iteratively revised and added to during the early stages of observation. After defining the specifics of the coding scheme, two analysts should establish IRR before moving forward with the analysis, if establishing relability is a priority for the research team [5].

In addition to transferring the codes to study visitor interactions with *GrooveMachine*, we have also modified the *specifics* section to create codes for two other exhibits—*Sound Happening*, an exhibit for playful music-making with colorful bouncy balls and *Dive Trainer*, an exhibit where learners can control an embodied simulation of a WWII U-Boat (see Table 1). These exhibits both differed significantly from the *TuneTable* iterations. *Sound Happening* is an exhibit that involves full body interaction in which users move colorful balls around an interaction space. *Dive Trainer* is a more constrained interaction in which individuals use a steering wheel and buttons to control a simulated U505 submarine. Only one user physically engages with the *Dive Trainer* interface at a time, but in practice it ends up being a collaborative exhibit because other group members observe and offer directions, feedback, and commentary.

The APEX codes we developed for *Sound Happening* and *Dive Trainer* have been preliminarily applied in informal studies but have yet to be operationalized in a larger scale formal study. However, developing codes for these exhibits helped to establish the transferability of the APEX framework to other collaborative, embodied museum exhibits. In particular, developing codes for *Sound Happening* helped us to ensure the general definitions and keywords for each code were applicable to full-body embodied interfaces in addition to tangible tabletop interfaces like *TuneTable*. Adapting the coding scheme to *Dive Trainer* indicated that we had developed a framework that could be applied to collaborative, embodied interactions with exhibits with more constrained interactions.

The full APEX coding scheme is described below, summarized in Table 2, and detailed in the codebook included in the supplemental materials.

Table 2. Summary of APEX codes, broken down by the category of engagement - physical, social, intellectual, and emotional.

Physical	**Social**	**Intellectual**	**Emotional**
P_Isolated_Manipulation	S_Discord	I_Seeking_Knowledge	E_Positive_Emotion
P_Investigative_Manipulation	S_Harmony	I_Sharing_Knowledge	E_Neutral_Emotion
P_Integrated_Manipulation		I_Applying_Knowledge	E_Negative_Emotion
	S_Independent		
	S_Collaborative		
	S_Active_Passive		
	S_Equal_Partners		

5.2.6 Additional Resources. In addition to the video coding scheme that we present below, we have released a variety of resources online and in the supplemental materials. This set of resources is intended to help other researchers apply APEX to novel exhibits. These resources include a detailed manual for applying the APEX coding scheme to a novel exhibit, guidance on how to record and process video data, templates and instructions on how to format data for analysis, and R code used to calculate inter-rater reliability, run the APEX analysis, and generate visualizations like those shown in section *6 Results*. We also include some more recent

expansions we have made to the APEX toolkit, including a form for unobtrusive in-situ observation and a worksheet for setting a priori design goals using APEX (discussed in more detail in *7 Expanding the Toolkit).*

5.3 Physical Engagement

5.3.1 Overview. We define *physical engagement* as the ways in which visitors interact with an exhibit by engaging in hands-on manipulation of the tangible or embodied aspects of the installation. This is a little narrower than the definition in the original APE codebook, which also includes factors like where the visitors are standing in relation to the exhibit [20]. The APEX framework could be expanded in the future to include this type of information, but for now we chose to focus on this narrower scope because 1) our data was not particularly illuminating in this respect, since visitors primarily stood in one place during their interactions and 2) we wanted the coding scheme to focus on visitors' engagement with the learning goals of the exhibit, which are materialized in visitors' direct interaction with the exhibit. We do not believe this hinders the transferability of the framework to other embodied exhibits, as the stages of physical engagement we describe below have been able to represent a wide range of interactions with full-body exhibits like *Sound Happening* (e.g. participants moving from kicking a ball to test out the exhibit to purposefully coordinating their movements in an effort to make different sounds). In addition, the flexibility in the "specifics" section of the coding scheme allows research teams to define what physical engagement with their particular exhibit looks like. However, we note that existing techniques for studying visitor position in museums could be used to supplement APEX data if the research team is specifically interested in tracking where visitors are located with respect to each other and the exhibit (e.g. [53,69]).

Our final codebook consists of three categories that reflect stages of physical engagement– *isolated manipulation*, *investigative manipulation*, and *integrated manipulation*. In the case of *Blockhead* and *GrooveMachine*, opportunities for participants to engage with a variety of computational concepts increase as participants move from *isolated* to *investigative* to *integrated* manipulation. This is echoed in VEF, where Barriault and Pearson characterize the physical aspect of progression from *initiation → transition → breakthrough* as *doing the activity once → repeating the activity → engaging in concentrated experimentation and exploration* [6]. We have written in more detail about the different physical engagement trajectories visitors may take while interacting with an exhibit in [47].

5.3.2 Coding Scheme. Isolated manipulation refers to moments when participants are physically interacting with the exhibit, initially testing/touching/trying out components in isolation to determine how individual pieces work. Keywords/phrases include: *trying out individual components*, *potentially unsure*, *disconnected*, *haphazard*, *isolated*. Some examples include placing a new sound block on the table (*Blockhead*) or randomly pressing arcade buttons (*GrooveMachine*).

Investigative manipulation refers to moments when participants are beginning to explore how the exhibit works by testing out two or more components together, but are not yet fluidly integrating multiple components in a complex sequence of actions. Participants are investigating the relationship among multiple pieces. Keywords/phrases include: *coordinating multiple elements*, *investigating relationships*, *testing*. Examples include making connections between sound blocks (*Blockhead*) or connecting a modifier block methodically to a sample block or to another modifier block (*GrooveMachine*).

Integrated manipulation refers to moments when participants are fluidly integrating multiple components in a complex sequence of actions. Participants are no longer investigating/testing and are now engaging in goal-oriented connection of multiple exhibit components. Keywords/phrases include: *executing complex actions*, *goal-oriented*, *fluid interaction*, *expressive*, *composing*, *confident*, *advanced components*, *integrating multiple components*. Examples include adding a sound or function block to a sound + function chain on the table (*Blockhead*), or using the joystick/buttons methodically after or while making a composition (*GrooveMachine*).

5.4 Intellectual Engagement

5.4.1 Overview. Humphrey et al.'s APE framework defines *intellectual engagement* as dealing with "the connections visitors make to existing knowledge during their interaction, the conceptual understandings [they gain], and the questions they have" [20]. APE codes for intellectual engagement were very exhibit specific. We use the same definition of intellectual engagement as the original APE framework, but break down intellectual engagement into three more general-purpose categories that were not in the original APE framework: *seeking knowledge*, *sharing knowledge*, and *applying knowledge*. Since APEX is a tool for in-situ observation and video analysis, this definition of intellectual engagement does not take into account visitors' prior knowledge (unless the visitor references their prior knowledge during the interaction). Other tools like personal meaning maps [23] may be more well-suited for assessing visitors' prior knowledge in museums.

Several existing tools for conversation analysis have been used in prior work to assess aspects of intellectual engagement (e.g. [1,61]). Our approach focuses on using observable verbal indicators to identify intellectual stages of engagement, inspired by VEF. VEF groups all verbal intellectual engagement (*seeking* and *sharing* information as well as aspects of *applying* knowledge like relating the exhibit to a prior experience) into *breakthrough*, the highest level of engagement [6]. Since APEX is targeted at active prolonged engagement exhibits, we hope to see progression in visitors' verbal intellectual engagement. To illuminate visitors' progress, we take VEF's description of intellectual engagement a step further and examine in a more fine-grained way what is happening within the *breakthrough* stage of visitor engagement. We suggest that the progression from *seeking* to *sharing* to *applying* knowledge is indicative that the participant is engaging with the exhibit and embedded concepts on a more complex level.

5.4.2 Coding Scheme. Seeking knowledge reflects moments when participants are seeking more knowledge about the experience. *Seeking knowledge* encompasses two types of behaviors:

1. *Asking Questions:* This involves asking questions about how aspects of the exhibit work (to oneself or others) (e.g. "What does this do?") and/or asking questions that promote curiosity or further engagement with the system. (e.g. "How do you think that works?"; "Do you want to try that?") This does not include extra-diegetic information or questions. Analysts should focus more on the inquisitive intent, and less on the grammatical form (e.g. "I wonder what this does" is asking an implicit question, even though it is a statement).
2. *Confusion:* This involves a visitor expressing confusion about an aspect of the system that violated their expectations (e.g. "Hmm…?"; "What just happened?").

Keywords/phrases include seeking more knowledge; questions/confusion about how the exhibit works and about what one might learn from the exhibit.

Sharing knowledge reflects moments when participants are sharing knowledge about the experience. Sharing knowledge encompasses two types of behavior:

1. *Voicing Observations:* This involves verbalizing what is happening in the exhibit (including reading signage or other information about the exhibit), or noticing that something is occurring (e.g. "It made a sound"; "This block has lines going this way and this one has lines going the other way"). This includes aesthetic observations (e.g. "This is cool"), but not aesthetic decisions (these fall under *applying knowledge*—e.g. "Use that block because that looks/sounds nice"). This does not include observations about unrelated things (e.g. "It's getting late") or observations about the state of other participants (e.g. "Oh you are just playing").
2. *Explaining:* This involves a visitor offering an explanation or hypothesis for how they think the system works, even if it is incorrect; explaining 'why' something is happening (e.g. "It might be happening because..."; "It probably reads all of it when you hit play").

Keywords/phrases include voicing observations about the exhibit or what they are learning through the exhibit; explaining/hypothesizing about how the system works or about the underlying concepts.

Applying knowledge describes moments when participants are applying their knowledge of the experience by planning or directing action. Applying knowledge encompasses a variety of behaviors, which are described below.

1. *Proposed Solutions:* The participant verbally proposes a solution to a problem they are trying to solve. This follows a violation of the user's expectations of how the system works (e.g. "What if we put this at the end?"; "Maybe we should move this").
2. *Planning:* The participant verbally proposes a goal or plan for the group. This should involve more than one step, be goal-driven, and move the group to a place where someone can conduct (e.g. "Let's see what they sound like individually"; "Let's add in a backbeat").
3. *Conducting:* The participant tells or suggests to another participant how to contribute to the composition; a command. This does not include proposing a goal for the group, but is rather a singular, action-driven process (e.g. "Play!"; "Move that there")
4. *Aesthetic Decisions:* The participant chooses to incorporate elements they like, discarding or removing elements of the activity they find displeasing (e.g. "Use that block because it sounds good"). This does not include aesthetic opinions that do not result in a decision (e.g. "That looks/sounds nice!")
5. *Prior Knowledge:* The participant explicitly, verbally relates the exhibit to other experiences in school, life, exhibits, etc. (e.g. "This reminds me of programming!"; "This sounds like the theme from Star Wars!")

Keywords/phrases include: Proposing a solution; making a plan for the group; telling another participant how to contribute; making choices based on aesthetics; applying prior knowledge.

5.5 Social Engagement

5.5.1 Overview. Humphrey et al. define *social engagement* as having to do with "the many ways in which visitors influence each other's experiences at exhibits" [20]. VEF did not explicitly address social engagement [6]. The original APE framework created two five-point scales for coding for social engagement: *independence* vs. *working collaboratively* and *harmony* vs. *conflict* [20]. The original APE framework analyzes the group's interaction as a whole (e.g. after the group interacted with the exhibit for several minutes), whereas APEX evaluates interaction for each 10 second interval. We found in practice that there was not a meaningful difference in a 10

second time interval between, for example, level 3 ("some conflict and some harmony") and level 4 ("more harmony than conflict"). As a result, The APEX framework reduces the five-point scales for *independence* vs. *collaboration* and *harmony* vs. *conflict* into two-point binary scales that are more well-suited for short units of analysis.

Our thematic analysis also revealed an additional perspective to consider when coding for social engagement. We added *active/passive* vs. *equal partners* to the framework based on our thematic analysis, which revealed a variety of teaching and leading/following social behaviors that were not captured by the APE framework's existing two scales. Each of the three binary scales (*independence vs. collaboration*, *harmony vs. conflict*, *active/passive vs. equal partners*) is described in more detail below. Two of the scales, *independence vs. collaboration* and *harmony vs. discord*, may be viewed as dominant and nondominant pairs—that is, the nondominant code will only be applied in the absence of any indicators of the dominant code. *Independence* and *harmony* are nondominant; *collaboration* and *discord* are dominant.

5.5.2 Coding Scheme. We define *discord* as a social conflict such as a disagreement, an interruption, or a disturbance to others' play—a break in the harmony. Although it does not necessarily have to be associated with a negative emotional response, it should in some way disrupt the flow of the play experience. Keywords include: *conflict*, *disruption*, *controversy.*

Some examples of discord include (c.f. [46] for more detail):

- *Conflicting Goals:* group members hold differing creative goals and conflict arises from deciding how to proceed
- *Opposing Hypotheses:* group members hold differing ideas of how the exhibit works and conflict arises from figuring out which one is correct
- *Taking Control:* one or more group members attempt to take control of the actions happening on the table and/or take over the work of others
- *Limited Space/Materials:* group members fight over scarce resources (such as tangibles or space at the exhibit) (e.g. "Hey that's mine!", fighting over a block)
- *Disruptive Distraction:* discord unrelated to the exhibit; often small children drawing attention away from the play experience.

We define *harmony* (the nondominant code) as working together in the absence of social conflict (not necessarily working together joyfully, per [20]).

We say an interaction is *collaborative* if at least two members of the group are collaborating (i.e. actively working towards a constructive, shared goal). Collaboration can be physical (e.g. working together on the same task) or verbal (e.g. directing or planning together). Keywords include: *working together*, *sharing space or tools*, and *shared planning.* Examples include handing another group member a sample/modifier block (*GrooveMachine)* or giving instructions for the modification of a chain (*Blockhead*: "Try connecting it like this").

An interaction is *independent* (the nondominant code) if no one in the group is working collaboratively; no indicators of collaboration are present. Keywords include: *working alone*, *parallel play*, *individual play.*

There is an *active/passive* relationship between the group members when some members in the group take on an active role, and some members in the group take on a passive role. We define *active* as teaching or directing/suggesting the action (e.g. explaining/narrating the experience, using facilitating language like "What do you think you should do?"; or using conducting language like "put that there"). We define *passive* as members that are listening/observing/doing what they are told, or simply failing to take part in an active role. Keywords include: *leader and follower dynamic*, *members contributing unequally.* Some examples

of active-passive group dynamics are listed below. We drew on existing research in the HCI community to define this list of social roles that people take on during co-creative interactions [45,48,57].

- *Teacher-Apprentice:* teacher is explaining exhibit, apprentice is listening and/or asking questions of teacher
- *Facilitator-Follower:* facilitator is asking guiding questions, follower is looking for answers
- *Leader-Follower:* leader is setting a course for the group, follower is doing what is suggested
- *Taskmaster-Worker:* taskmaster is giving specific instructions, worker is carrying them out
- *Actor-Observer:* actor is interacting with exhibit, observer is watching interaction or exhibit
- *Actor-Commentator*: actor is interacting with exhibit, commentator is remarking on exhibit but not interacting

An *equal partners* dynamic is any dynamic that is not active/passive. Either no one in the group has taken on an active role *or* everyone in the group has taken on an active role. This may involve turn-taking between all members within a segment. Keywords include *equal contribution*, *no dominant member*, *all guiding action equally*. Some examples of equal partners group dynamics include:

- *Actor-Actor:* all group members are interacting with exhibit
- *Commentator-Commentator*: all group members are remarking on exhibit but not interacting with it
- *Taskmaster-Taskmaster*: all group members are giving specific instructions
- *Observer-Commentator*: observer(s) is/are watching exhibit, commentator(s) is/are remarking on exhibit
- *Teacher-Facilitator:* teacher(s) is/are explaining exhibit, facilitator(s) is/are asking guiding questions or suggesting a course of action
- *Leader-Taskmaster:* leader(s) is/are setting a course for the group, taskmaster(s) is/are giving specific instructions

5.6 Emotional Engagement

5.6.1 Overview. The original APE framework defines *emotional engagement* as having to do with the "nature and intensity of the affect exhibited by visitors during the engagement and immediately after" [73]. The nature of the emotional engagement may be *positive*, *negative*, or *neutral*. We use the same three codes for emotional engagement as the original framework and supplement them with concrete examples of what these types of engagement look like in practice. We classify positive and negative emotion as dominant codes and neutral emotion as a nondominant code—i.e. it is only applied in the absence of any indicators of positive/negative emotion. Positive emotional expression is coded as a *transition* behavior in VEF [6].

5.6.2 Coding Scheme. Positive emotional engagement is defined as a positive expression towards the experience. This includes positive reactions to the experience (e.g. "That's cool"; "Wow!"; "I like that"), positive reactions to others, within the context of experience (e.g. "Great idea!"; "You're so smart!"), positive body language/noises (e.g. dancing, clapping, high fives), and

positive laughing (e.g. amused laughter or laughter indicating enjoyment/joy). Keywords include *happy*, *joyful*, *supportive*, and *positive*.

Negative emotional engagement is defined as negative expression towards the experience. This includes negative reactions to the experience (e.g. Wow that's lame."; "That sounds bad"), negative reactions to others, within the context of the experience (e.g. yelling something; "Yours is weird"; "She broke it!"), and negative body language/noises (e.g. stomping; hand swatting; crying; wailing). This does not include conducting action (i.e. telling others what to do—e.g. "Stop that"), as conducting is more directional than emotional. Keywords include: *negative*, *unhappy*, *sad*, *angry*, *disappointed*, *disgruntled*, and *mean*.

Neutral emotional engagement is the nondominant code and is defined as the absence of a positive/negative emotional response. In practice, we found that the neutrality of certain vocalizations was a bit ambiguous, so we provide some clarifying points in the following paragraph. Keywords/phrases include: *neutral*, *not positive or negative*.

There may be *energy* in a neutral statement even if there is not *emotion* (e.g. "Look at that!" said with energy is not necessarily positive or negative). We also classify the following expressions as *neutral*: emotional responses that do not relate to the exhibit, interjectional "ums" and "ahs," conducting or telling others what to do, expressing curiosity, making observations, apologizing, and laughter that is not clearly expressing a positive emotion.

6 RESULTS

We applied this coding scheme to studying participant interactions with both *Blockhead* and *GrooveMachine*. For each exhibit, we present a case study of how we used APEX to conduct 1) a macro-level analysis of findings/insights about the exhibit and 2) a micro-level analysis of the learning trajectory of a single interaction group.[2] This is intended to demonstrate 1) how to operationalize APEX in practice and 2) the types of insights that the APEX coding scheme can provide to researchers and designers.

The case studies of *Blockhead* and *GrooveMachine* that we present here are not intended to be directly compared with one another. We did not conduct user studies with controlled variables and the study setups for the two exhibits differed significantly (see *3 The Exhibits* for details—most notably, *Blockhead* was installed in a museum classroom environment while *GrooveMachine* was installed on the museum floor). Rather, these two case studies are intended to demonstrate the types of insights that APEX can provide about individual exhibits. We discuss how APEX might be used in the future to more directly compare exhibit iterations in *7 Expanding the Toolkit*.

6.1 Inter-Rater Reliability

We calculated an inter-rater reliability (IRR) score for each type of engagement using a subset of the video data after iteratively refining our themes using reliability as a tool. Two coders were used to establish IRR for each category, although coder pairs differed from category to category.

2 We additionally refer the interested reader to two papers we have previously published that present detailed analyses of specific sub-components of the APEX framework (e.g. focusing specifically on physical engagement or discord/harmony) [46,47]. These papers illustrate the depth of insights that APEX can provide, depending on the questions the research team is interested in.

Coder pairs analyzed a subset of the video data for each exhibit in order to establish IRR. The rest of the video data was coded by a single analyst.

We use Gwet's AC1 [29] statistic to calculate IRR scores, due to a recognized issue with Cohen's Kappa when it is calculated for data in which certain events (e.g. discord, positive and negative emotion) are rare [75]. The AC1 statistic is an alternative to Cohen's Kappa that corrects for this issue while still accounting for chance agreement [29]. IRR scores for each of the categories of engagement are reported in Table 3. All scores are classified as either substantial agreement (.61 to .80) or almost perfect agreement (.81 to 1.00) according to [78].

Table 3. Inter-rater reliability Gwet AC1 scores for all coding categories for both exhibits (social codes are broken down by sub-category)

Code Category	**Gwet's AC1 - *Blockhead***	**Gwet's AC1 - GrooveMachine**
Intellectual	.69	.71
Social		
Active-Passive/Equal Partners	.65	.83
Discord/Harmony	.93	.99
Independent/Collaborative	.81	.79
Physical	.92	.82
Emotional	.85	.95

6.2 Blockhead

6.2.1 Macro-Level Analysis. In the macro-level analysis, we look at the time spent by visitors in each category of engagement across all 31 participant groups (Figure 3). This is similar to the type of analysis that the VEF and APE support, but goes a step further by looking not just at *whether or not* participants engaged with the exhibit in a certain way, but also *how much time* they engaged in that way for. This high-level view has the advantage of revealing common patterns across multiple groups of visitors. We see that visitors engaged in integrated physical engagement with *Blockhead* for 71% of the time. This indicates that most participant groups advanced to using both sound and function blocks and were able to fluidly use the blocks to create compositions. A further analysis reveals that on average, groups took about 2 minutes to reach a period of sustained (i.e. one or more minutes) of integrated manipulation. The boxplots in Figure 3 show that there was some variation amongst participant groups as to the amount of time they spent in each stage of physical engagement, but that no group spent more than 45% of the time in isolated engagement (the "lowest" stage).

Intellectual engagement was more evenly split across engagement categories. Visitors engaged in sharing knowledge 55% of the time, seeking knowledge 42% and applying knowledge 43%. This finding, combined with visualizations of individual group interactions like the one shown in Figure 4, suggests that rather than progressing directly from seeking to sharing to applying knowledge, visitors fluidly shifted between the three types of intellectual engagement throughout their interaction. The box-plot visualization (Figure 3) shows that there was also wide variation amongst visitor groups in terms of how much time they spent in each stage of intellectual engagement.

The social engagement analysis reveals that participant groups engaging with *Blockhead* spent most of their time in an *active/passive* dynamic. This contrasted with our original expectations, since we had designed the exhibit to foster an environment in which visitors were co-creating as equals. However, the active/passive dynamics usually emerged as parents

prompted their kids to interact or asked them guiding questions, or as one member of the group took charge and directed others' movements (this sort of parental guidance/scaffolding is common in museums [70]). This is not necessarily detrimental to learning; Vygotsky's theory of the zone of proximal development argues that a social learning partner like a teacher, parent, or more advanced peer can help to push an individual to learn more than they would on their own [76].

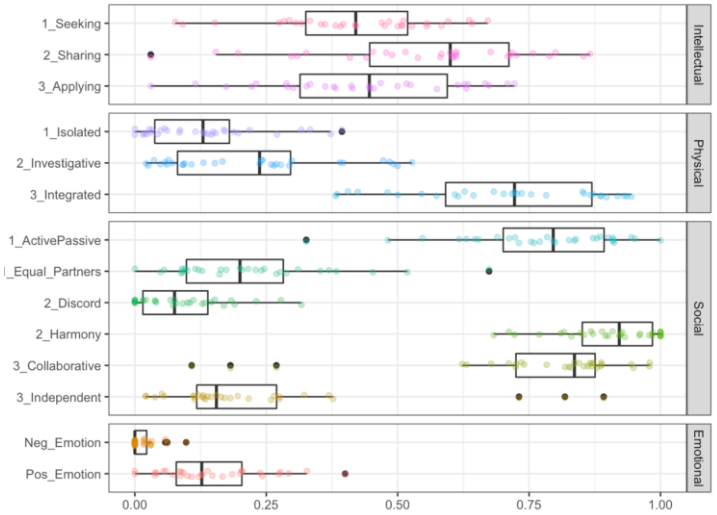


Figure 3. Box plots depicting percentage of time participants spent in each APEX engagement category for *Blockhead*.

We also found that visitor interactions were harmonious for the vast majority of the time. Only 9% of interactions had moments of discord. Although moments of discord are rare, they can be significant. Moments of discord can indicate setbacks that the group faces and overcomes and can highlight moments when participants are engaged in a "flow" state and do not wish to be interrupted [46].

We also found that participants were engaged in *collaborative* group interactions the majority of the time. This is a positive outcome for this particular exhibit as we had sought to promote group collaboration in the design of *Blockhead*. *Individual* work still occurred but was less frequent.

Finally, the analysis shows that instances of emotion were rare, but that *positive* emotional reactions were more frequent than *negative* emotional reactions. This is also a positive outcome, as positive expressions of emotion can indicate that a visitor group is comfortable interacting with an exhibit and motivated/eager to continue engaging [6].

6.2.2 Micro-Level Analysis. We take a closer look here at one particular participant group that interacted with *Blockhead* as a case study to aid in understanding the types of insights that analysis using the APEX framework can provide at a micro-level. This particular group was composed of three participants—a mother and her son and daughter. Their interaction with the table lasted a little less than ten minutes. The visualization in Figure 4 depicts the different codes that participants exhibited during each ten second segment of the interaction. Since multiple intellectual and physical codes could be applied to each time segment, those are shaded according to the highest level of engagement that was occurring at that time.

Participants displayed neutral affect for the majority of the time (60%). There were 22 spikes of positive affect (indicated by the green triangles in Figure 4). Participants expressed positive emotion in a variety of ways, including dancing, lots of laughter (especially when they accidentally created an infinite loop), and getting excited about discovering what blocks did. The existence of 22 positive reactions during the course of an approximately ten minute interaction indicates the potential of the exhibit to improve participants' impressions of computing.

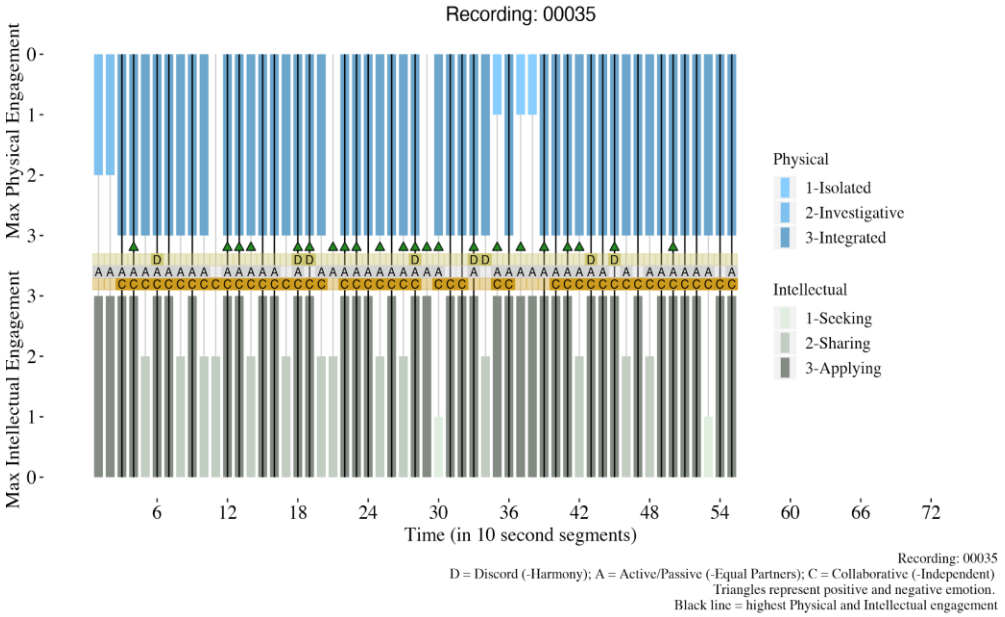


Figure 4. Visualization of the APEX codes applied to each 10 second segment of a single group's interaction with *Blockhead*, revealing the physical, social, emotional, and intellectual interaction trajectory for the group.

Social interaction was mostly harmonious (85%) with eight occurrences of discord. Most instances of discord were caused by limited space/materials (e.g. girl stealing some blocks from her brother) or conflicting goals (e.g. daughter trying to remove a loop block, but the mom says she wants her to leave it to see what it does). The group spent the majority of the interaction time (84%) in active/passive mode, with the mother taking the lead in the interaction by instructing her kids and guiding their compositions. The group also worked collaboratively for the majority of the interaction time (84%).

The group appeared to quickly grasp the interaction and begin to engage in fluid exploration/composition. The participant group starts their interaction by connecting multiple sample blocks after a brief introduction from a facilitator. As soon as the group figures out what a function block is, they quickly ramp up to building a pretty complex composition comprised of many sample and function blocks. The group remains in this stage of integrated manipulation for 85% of the interaction. The group fluidly moves between different types of intellectual engagement, alternating pretty evenly between *seeking* (29%) *sharing* (39%) and *applying* (39%) knowledge throughout the interaction.

6.3 GrooveMachine

6.3.1 Macro-Level Analysis. The macro-level analysis for *GrooveMachine* (Figure 5) reveals different patterns of engagement than what we saw in *Blockhead*. The participants did not engage in a lot of intellectual engagement overall (9% seeking, 9% sharing, and 9% applying). The box plots (Figure 5) indicate that there were not vast differences in these numbers across participant groups, although some groups engaged in *sharing knowledge* for a slightly larger portion of the interaction. The APEX framework exclusively focuses on *observable, social* behaviors and does not capture instances of more silent, introspective learning. Therefore, this finding does not necessarily mean that visitors did not find the exhibit engaging or that they did not learn from it. It does however indicate that the exhibit did not foster a lot of group dialogue/visible "learning talk." Learners also did not observably progress from *seeking* to *sharing* to *applying* knowledge.

The physical engagement analysis indicates that participants were physically engaging with the exhibit for the vast majority of the interaction time. Interaction hold times were also quite long (5:34min), indicating the exhibit was engaging. However, participants primarily engaged in *isolated manipulation* (90% of the time). The box plots (Figure 5) indicate that this was fairly consistent across visitor groups.

Although more visitors engaged in an *active/passive* social dynamic with *GrooveMachine*, the box plots (Figure 5) indicate that social dynamics varied quite a bit between participant groups and many engaged as *equal partners* for a significant portion of the interaction time. This may be a strong point of the exhibit. *GrooveMachine* was designed with four different interaction stations, which was intended to allow learners to have individual ownership over their section of the table even while contributing to a group composition. This finding indicates that the designated table sections might have fostered an "equal partners" engagement dynamic, allowing multiple people to work on the table simultaneously without one person explicitly taking charge. This might allow for more people to actively co-construct knowledge simultaneously, rather than one person constructing knowledge and others following along or participating peripherally.

Visitors were engaged in *harmonious* interaction nearly the entire time, with very few instances of *discord*. This could also be a result of the individual working space—visitors may not have had territorial conflicts because they had plenty of space in which to work separately.

This exhibit fostered more *independent* work (59%) than *collaborative*. Although the design team wanted to create a space for individual composition, the primary objective of the exhibit was to foster a co-creative experience, so the dominance of individual work is not ideal. This may be related to participants working alone within their own sectors of the table rather than all working together on the same composition. This could also explain why visitors were not engaging in very much collaborative dialogue, since they were playing individually. Future design iterations might explore how to achieve a better balance between individual composition and group collaboration/dialogue.

Finally, instances of emotional reaction were rare, with just a few instances of *positive* emotional reaction and very few instances of a *negative* reaction. The low amount of expressed emotion could be related to the lack of group dialogue, since APEX is focused on clearly observable markers of emotion that are often shared more when in social dialogue with others.

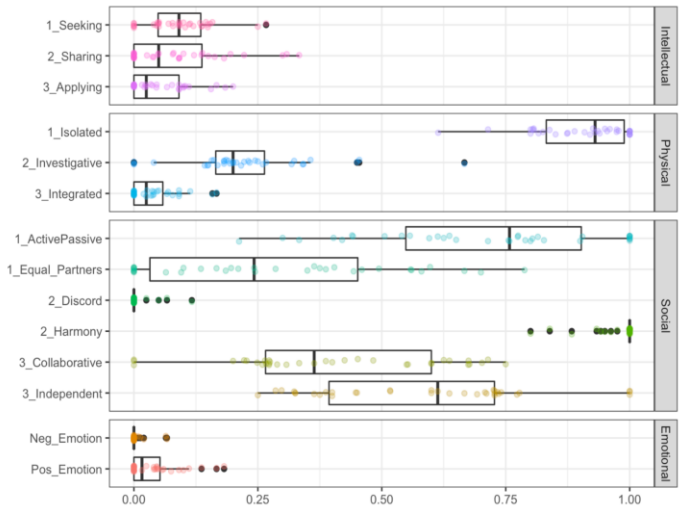


Figure 5. Box plots depicting percentage of time participants spent in each APEX engagement category for *GrooveMachine*.

6.3.2 Micro-Level Analysis. The group we describe here as a case study for *GrooveMachine* consisted of four total participants—a mother, father, and two daughters (one very young and one older). This interaction lasted about 7 and a half minutes. The mother and two daughters were most active in this timeframe and the father interacted for the first 2 minutes then became an inactive participant (i.e. watching but not commenting or interacting).

Within this group, the participants displayed neutral affect for the majority of the time (90%) and positive affect for 10% of the time. An example of a positive effect was when the mother expressed how she "likes the violin sounds" to her eldest daughter. Overall there were only 4 spikes of positive affect over the 7.5 minute long interaction and they happened after the first few minutes of interaction which may indicate progression in the group's engagement and motivation [6].

The group spent most of their time in harmonious interactions (97%) with just two instances of discord due to conflicts over control of the table between the mom and eldest daughter. Throughout the interaction, the mom was in control of the group as they worked in active/passive mode for 60% of the time. The group was collaborative for 67% of the time, and the visualization in Figure 6 suggests that most of the collaborative time was also active/passive. This group worked mostly collaboratively at the beginning of the interaction to figure out how things worked as a group and then there were some instances of independent work in which members discovered something on their own and brought it back to the group as a whole. This kind of interaction points to a fluid transition between individual exploration and social learning.

The group largely remained in isolated manipulation during the first 5 minutes of their interaction. Towards the last few minutes of engagement, the participants began to explore the connected sound samples by creating compositions of the blocks to make sounds that they enjoyed. For example, the mother expressed how she liked the violin sound so they worked to incorporate that into the composition. Although this group spent a lot of time in the isolated manipulation stage, their later advancement to investigative manipulation indicated that they were learning and developing an understanding of the exhibit and embedded concepts.

Lastly, out of the 7.5 minute interaction, the group spends 43% of the time sharing knowledge with each other. For example, the mom explained to her youngest daughter that the piece she was playing with "is an outside piece." The group engaged in *seeking knowledge* 20% of the time and *applying knowledge* 17% of the time. The remaining time was coded *no code*, meaning the conversations did not fall under any of the three intellectual categories. This group spent the first half of the time seeking and sharing knowledge, and their first instance of applying knowledge occurred around the four minute mark. Here, the mother began applying information that she learned by actively pushing the blocks together. This shows that the group progressed from trying to understand how everything worked to then applying what they had learned.

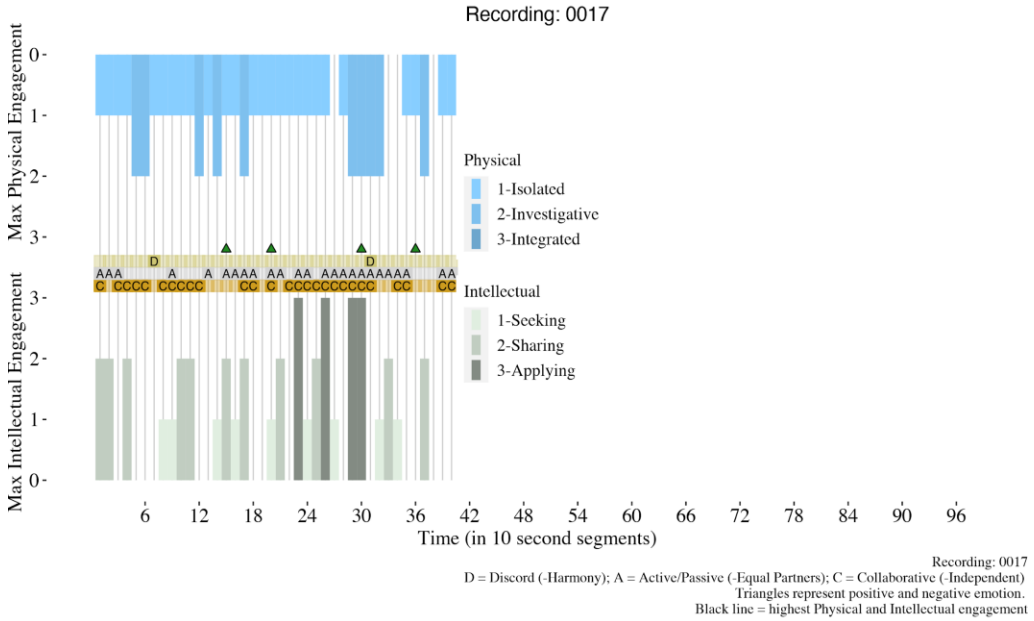
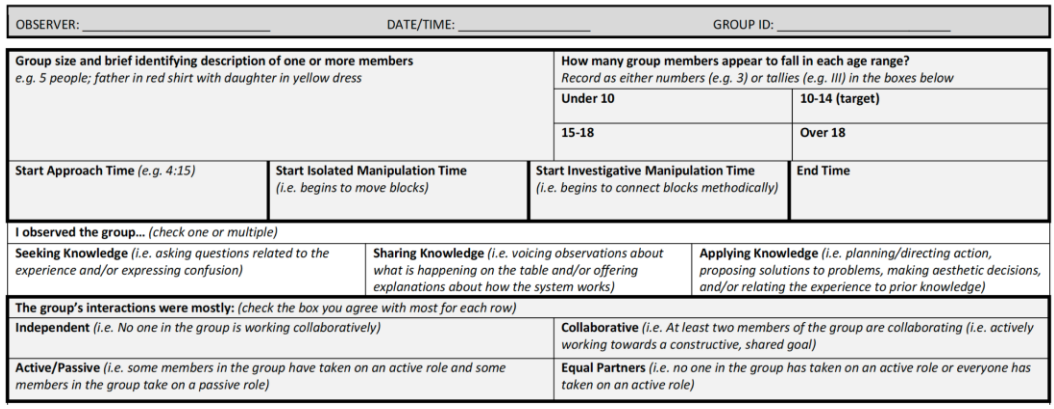


Figure 6. Visualization of the APEX codes applied to each 10 second segment of a single group's interaction with *GrooveMachine*, revealing the physical, social, emotional, and intellectual interaction trajectory for the group.

7 EXPANDING THE TOOLKIT

We originally developed the APEX coding scheme as a framework to guide qualitative video analysis. We have recently expanded the APEX toolkit to include instruments for in-situ observation—for researchers and practitioners who would like to use APEX but do not have the time or resources to conduct a lengthy video analysis—and a worksheet for design teams to use to set a priori goals against which to evaluate their exhibits. This section describes these additional instruments in more detail.

7.1 Observation Form



OBSERVER: _____ DATE/TIME: _____ GROUP ID: _____

Group size and brief identifying description of one or more members *e.g. 5 people; father in red shirt with daughter in yellow dress*	How many group members appear to fall in each age range? *Record as either numbers (e.g. 3) or tallies (e.g. III) in the boxes below*	
	Under 10	**10-14 (target)**
	15-18	**Over 18**

Start Approach Time *(e.g. 4:15)*	Start Isolated Manipulation Time *(i.e. begins to move blocks)*	Start Investigative Manipulation Time *(i.e. begins to connect blocks methodically)*	End Time

I observed the group... *(check one or multiple)*

Seeking Knowledge *(i.e. asking questions related to the experience and/or expressing confusion)*	Sharing Knowledge *(i.e. voicing observations about what is happening on the table and/or offering explanations about how the system works)*	Applying Knowledge *(i.e. planning/directing action, proposing solutions to problems, making aesthetic decisions, and/or relating the experience to prior knowledge)*

The group's interactions were mostly: *(check the box you agree with most for each row)*

Independent *(i.e. No one in the group is working collaboratively)*	Collaborative *(i.e. At least two members of the group are collaborating (i.e. actively working towards a constructive, shared goal)*
Active/Passive *(i.e. some members in the group have taken on an active role and some members in the group take on a passive role)*	Equal Partners *(i.e. no one in the group has taken on an active role or everyone has taken on an active role)*

Figure 7. Snippet showing a portion of the physical APEX observation form. The full form (in physical and virtual formats) is attached in the supplemental documents.

Qualitative video analysis is commonly used in the research community, but for museum practitioners or design teams looking to rapidly iterate, the process of collecting and coding video data can be too labor-intensive and time-consuming. For practitioners interested in understanding participant engagement who do not have the time or resources to devote to a fine-grained video analysis procedure, we provide forms for live, in-situ observation using the APEX framework. While the data collected with these forms does not provide the depth and detail of a full APEX analysis, they should give a high-level understanding of participants' social, intellectual, physical, and emotional engagement with a much lighter time/resource commitment. Observations can also be used to provide more rapid/instantaneous data to supplement the lengthy video analysis process (e.g. to aid in rapid iteration without losing the detail of a full analysis). See Table 4 for a summary of time/resource requirements for the different components of the APEX toolkit. We have tested and refined the observation form in informal micro-studies with a variety of exhibits–including *TuneTable*, *Sound Happening* , and *Dive Trainer* (described in *5.2.5 Transferring Coding Scheme to New Exhibits*). However, we have yet to use the form in a larger-scale formal study.

We provide both a physical form and a virtual form in the supplemental materials (a snippet of the physical form is shown in Figure 7). The virtual form makes it easier to simultaneously observe multiple or overlapping groups, but using the physical form may be less obtrusive depending on the exhibit setup. Both forms are based on the APEX framework and ask questions about group composition, timestamps for visitors transitions between stages of engagement, and high-level observation of whether or not certain APEX codes occurred.

7.2 Design Worksheet

In our analysis of *Blockhead* and *GrooveMachine*, we use APEX as an exploratory tool to understand the myriad ways in which visitors engaged with and interpreted the exhibits. We did this in the spirit of facilitating multiple design interpretations [67] and free-choice learning [23], in which different visitor groups may have different learning outcomes or interpretations of the same exhibit. This is a perfectly valid way to use the instrument and can lead to emergent insights about participant engagement that researchers may not have anticipated.

Table 4. Summary of time/resource estimates for different components of the APEX toolkit. Different aspects of the toolkit can be used depending on the team's needs/resources.

Toolkit Component	Resources Needed	Time Estimate
Video analysis	1-2 video cameras Tripod Overhead camera mount (specifics depend on space) Microphone(s) and mount(s) Consent sign At least 1 person to monitor video/exhibit At least 2 video analysts 1 person to prepare data (e.g. stitch together video streams + audio) Video transcription service	Depending on visitor density, 3 hours to 2 days of video recording 1 week for video preparation 1-3 months for video analysis, depending on quantity of video and number of analysts
Observation form (virtual)	1-2 computers 1-2 observers	Depending on visitor density, 3 hours to 2 days of live observation ~1 day for analysis
Observation form (physical	1-2 observers 1-2 stopwatches, clipboards, pens ~100 printed copies of form	Depending on visitor density, 3 hours to 2 days of live observation ~1 day for data entry ~1 day for analysis
Design worksheet	Design team Printed copies of worksheet Pens	30 min – 1 hour to fill out worksheet

However, some designers/researchers may want to take a less open-ended approach and assess whether specific a priori design goals were met in order to determine the "success" of an exhibit or to directly compare exhibit iterations. We were not able to directly compare the *Blockhead* and *GrooveMachine* exhibit iterations both because of differing study setups, but perhaps more importantly because we were still developing APEX at the time and had not set clear a-priori design goals related to the APEX dimensions of engagement. For example, visitors at *GrooveMachine* engaged in less verbal intellectual dialogue than *Blockhead*. This seems initially like it would be an indicator that *Blockhead* did a better job of facilitating learning and engagement, but researchers have found that some exhibits that prompted more silent reflection could lead to significant learning gains and that more "talk" is not always better [2]. We could have more clearly compared the *Blockhead* and *GrooveMachine* iterations if we had a "gold standard" metric to guide the design and evaluation process (e.g. "we want learners to engage in social dialogue that advances from seeking information to integrating knowledge" or "verbal dialogue is unimportant to us as long as visitors are progressing in physical engagement with the exhibit").

We have developed a preliminary instrument for designers based on the APEX framework that can be used to set a priori design goals and to directly assess whether or not the exhibit met these design goals. The instrument is a simple worksheet that prompts designers to indicate how much time their "ideal" user group would spend exhibiting each type of engagement in the

APEX framework (see Figure 8 for a snippet and the supplemental materials for the full worksheet). Designers are also prompted to circle the three types of engagement that are of highest priority to their team. Design teams can then compare this completed form to the results from the final APEX evaluation of their exhibit (Figure 8) in order to evaluate their exhibit according to the team's specific set of a priori goals. Free response questions on the worksheet prompt design teams to reflect on why their goals were or were not met. This will enable them to clearly assess whether their design goals were met in addition to highlighting surprising or unexpected results. This could potentially be useful for clearly communicating results with supervisors or funding agencies. We plan to use this tool in our future work when we are seeking to directly compare exhibit iterations or evaluate an exhibit against a "gold standard" benchmark. We also plan to continue to use APEX as a more open-ended/exploratory tool when we are looking to identify emergent patterns.

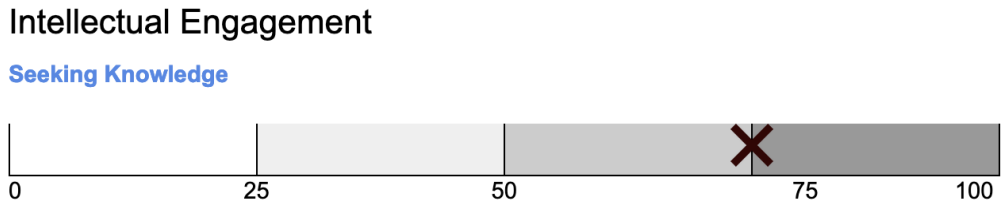


Figure 8. Snippet showing a portion of the APEX design worksheet. Here, designers can mark the desired percentage of total interaction time an "ideal" participant group would spend *seeking knowledge* in order to set a priori design goals. The worksheet can then be compared with results like those shown in Figure 4 and Figure 6. The full worksheet is attached in the supplemental documents.

8 TAKEAWAYS

The Results section illustrates that APEX enabled us to learn a lot about visitor engagement with both exhibit iterations— *Blockhead* and *GrooveMachine*—from both a macro-level view (i.e. across multiple participant groups) and from a micro-level view of detailed qualitative descriptions of each group's interaction trajectory.

While preliminary, these types of findings illustrate the depth of understanding that the APEX framework can provide that neither APE nor VEF offer on their own. APE might tell us that "most visitor groups had meaningful physical engagement with the exhibit"; VEF might tell us that "30% of visitors displayed breakthrough behaviors"; APEX goes beyond that and helps us to understand when participants reached a stage of engagement, whether they built up to that stage or jumped straight into it, and what indicators exist that participants reached that stage of engagement (e.g. Did they just reach physical integrated manipulation? Or was this also corroborated by social collaboration and intellectual application of knowledge?). APEX also allows us to understand this type of information on both an individual group level and a more macro-level view of all group interactions. These observations would also not have been made salient from other evaluation instruments commonly used in museums (e.g. [6,22,39,62]) or for understanding collaborative work (e.g. [59]).

Ideally, APEX results can aid research and design teams in reporting results from exhibits to relevant stakeholders. To make reporting easier, researchers may want to isolate specific elements of the plots and compare them to a priori design goals. For example, imagine an exhibit designed to challenge commonly held beliefs. At the outset, the design team has

indicated two important features: 1) they want participants to "apprehend" the exhibit fairly quickly (within the first 30 seconds) in order to experiment with the concepts; 2) the exhibit challenges a commonly held belief about a scientific phenomena; therefore, the designers expect participants to express discord, extremes of positive and negative emotion, and to move into physical:investigative and intellectual:sharing states after being in intellectual:applying in order to test the concept that challenges their commonly-help belief. The researchers may report plots that show these phenomena one at a time. The first plot might show apprehendibility by revealing only the physical and intellectual engagement elements. A second set of plots may focus on how designers feel the participants will respond to concepts that challenge their thinking: that is, a plot may show only social:discord and emotion and then overlay physical:investigative, intellectual:sharing, and intellectual:applying.

8 LIMITATIONS AND FUTURE WORK

We have currently applied the APEX framework to studying two iterations of an exhibit design—*Blockhead* and *GrooveMachine*. These were both exhibits focused on providing a collaborative, embodied computing learning experience. We assert that the process we went through to adapt the framework from *Blockhead* to *GrooveMachine* resulted in a set of codes and instructions for applying them that can be transferred readily to other collaborative, embodied exhibits that aim to foster active prolonged engagement, due to the distinctly different implementations of the two exhibits. We have preliminarily verified this by developing coding schemes for two other exhibits—*Sound Happening* and *Dive Trainer*—which differ from the *TuneTable* iterations in that they involve full-body interaction (*Sound Happening*) and fewer degrees of freedom/less physical collaboration (*Dive Trainer*). However, these coding schemes have not yet been operationalized in a formal study. Further study is needed to fully verify the transferability of the coding scheme.

The APEX framework is intended to provide an overarching view of how participants progress through stages of engagement. Other existing tools provide more in-depth analysis of particular aspects of participant engagement—for instance, frameworks for analyzing learning talk in conversation provide more insight into relevant discussion of content knowledge [61], techniques for tracking participant location could deepen researchers' understanding of physical engagement [53,69], and interviews or personal meaning maps could provide more insight into visitors' prior knowledge and expectations [23]. APEX can be used alongside or in addition to other analyses depending on the researcher/evaluators' needs.

Codes in each APEX category could potentially be expanded in future work depending on researchers' interests and resources. For example, researchers might collect visitor data using additional sources (e.g. eye tracking data, close-up video of facial expressions), which could allow for the addition of new emotional engagement codes. Researchers may also want to conduct additional data analysis depending on their interests. For example, visitors' progression through stages of engagement are currently more salient in the micro-analysis than in the macro-analysis visualization. In the micro-analysis, researchers can clearly see how the APEX stages of engagement for a single group change over time. The macro view alternatively allows researchers to summarize the amount of time groups spent in each stage on average (but does not present overarching statistics about patterns of progression, for example). Research teams could use the APEX data to generate such statistics in the future. In summary, the framework is adaptable and can be modified to suit a research group's specific needs.

One limitation is that the amount of data shown in the micro-visualization may be overwhelming and may provide plots that do not easily invite comparison. For example, a group that engages with an exhibit for one minute will have a different plot and therefore different experience than a group that engages for ten minutes. One approach is to normalize the plots to allow for comparison by showing what the group did in the first 25% of the interaction and in the last 25% of the interaction, for example. However, depending on the research team's questions, it may be more appropriate to group the various plots by feature and then analyze the groupings. That is, we may want to know what the group dynamics, size, age, etc. of the participants are for the 1-minute vs. the 4-minute group in order to better understand how the exhibit performs under different group dynamics. Similarly, the design worksheet outlines the expectations that designers have of participant engagement. We may then compare the expectations to actual participant engagement to better understand design decisions. Finally, instead of normalizing the plots, we may use the plots as they are to understand important phenomena. For example, most designers want groups to understand how the exhibit works quickly in order to enable them to experiment with the concepts of the exhibit (i.e. immediate apprehendability [35]). For this, we may analyze the first minute of engagement to determine whether and how long it takes participants to navigate through physical:isolation and intellectual:seeking in order to reach intellectual:applying.

9 CONCLUSION

In this paper, we present APEX, a framework that builds on both prior work and experimentally derived data to provide a coding scheme for qualitative analysis of visitors' physical, social, intellectual, and emotional engagement with collaborative, embodied museum exhibits that foster active prolonged engagement. We present two case studies where we apply APEX to understanding visitor interactions with different iterations of the *TuneTable* exhibit. These case studies illustrate that APEX illuminates findings at both a macro- and a micro-level that go beyond what previously developed evaluation instruments have been able to achieve. We additionally present an APEX observation form for lower-resource analysis and a preliminary tool that designers can use to set and evaluate a priori design goals using APEX, should they wish to use the framework in this way.

The framework, applied case studies, and preliminary design tool presented in this paper can be of use to designers, researchers, and educators studying collaborative, embodied, active prolonged engagement exhibits in informal learning spaces. The tools we provide can be used to facilitate a more informed, evidence-based iterative design cycle and to provide detailed insight into visitors' engagement and learning trajectories when interacting with exhibits. To facilitate the use of APEX by a wider audience, we provide a set of supplemental materials along with this paper submission (described in more detail in *5.2.6 Additional Resources*).

ACKNOWLEDGMENTS

We would like to thank all past and present members of the *TuneTable* evaluation and design teams, especially Jason Freeman, Astrid Bin, Anna Weisling, Anna Xambó, Gerard Roma, Hannah Guthrie, William Martin, Emily Bryans, Steven Blough, and Katlyn Voravong, without whom this research would not have been possible. We would also like to thank Aaron Price and the Museum of Science and Industry, Chicago for supporting and providing an audience for this research. This project is funded by the National Science Foundation (DRL #1612644).

REFERENCES

[1] Sue Allen. 2003. Looking for learning in visitor talk: A methodological exploration. In *Learning conversations in museums*. Routledge, 265–309.

[2] Sue Allen. 2004. Designs for learning: Studying science museum exhibits that do more than entertain. *Science Education* 88, 1: S17. Retrieved October 31, 2015 from http://xa.yimg.com/kq/groups/28001072/1514719950/name/Allen_Exploratorium.pdf

[3] Sue Allen, Patricia B Campbell, Lynn D Dierking, Barbara N Flagg, Alan J Friedman, Cecilia Garibay, and David A Ucko. 2008. Framework for evaluating impacts of informal science education projects. In *Report from a National Science Foundation Workshop. The National Science Foundation, Division of Research on Learning in Formal and Informal Settings*.

[4] Roya Jafari Amineh and Hanieh Davatgari Asl. 2015. Review of constructivism and social constructivism. *Journal of Social Sciences, Literature and Languages* 1, 1: 9–16.

[5] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The place of inter-rater reliability in qualitative research: An empirical study. *Sociology* 31, 3: 597–606.

[6] Chantal Barriault and David Pearson. 2010. Assessing Exhibits for Learning in Science Centers: A Practical Tool. *Visitor Studies* 13, 1: 90–106. https://doi.org/10.1080/10645571003618824

[7] Leslie Bedford. 2014. *The Art of Museum Exhibitions: How story and imagination create aesthetic experiences*. Left Coast Press.

[8] Ben Bengler and Nick Bryan-Kinns. 2014. In the wild: evaluating collaborative interactive musical experiences in public settings. In *Interactive experience in the digital age*. Springer, 169–186.

[9] Ben Bengler and Nick Bryan-Kinns. 2015. " I could play here for hours.."(thinks the visitor and leaves) Why People Disengage from Public Interactives. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 177–180.

[10] David Benyon. 2014. Spaces of interaction, places for experience. *Synthesis Lectures on Human-Centered Information* 7, 2: 1–129.

[11] S. M. Astrid Bin, Christina Bui, Benjamin Genchel, Kaushal Sali, Brian Magerko, and Jason Freeman. 2019. From the museum to the browser: Translating a music-driven exhibit from physical space to a web app. In *Proceedings of the International Web Audio Conference* (WAC '19), 24–29.

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2: 77–101.

[13] Harry Brignull and Yvonne Rogers. 2003. Enticing people to interact with large public displays in public spaces. In *Proceedings of INTERACT*, 17–24.

[14] Jerome S. Bruner. 1986. *Actual minds, possible worlds / Jerome Bruner*. Cambridge, Mass.: Harvard University Press, 1986. Retrieved from http://prx.library.gatech.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat01325a&AN=gatech.291392&site=eds-live&scope=site

[15] Leah Buechley, Mike Eisenberg, Jaime Catchen, and Ali Crockett. 2008. The LilyPad Arduino: using computational textiles to investigate engagement, aesthetics, and diversity in computer science education. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), 423–432. https://doi.org/10.1145/1357054.1357123

[16] Luigina Ciolfi, Gabriela Avram, Laura Maye, Nick Dulake, Mark T Marshall, Dick van Dijk, and Fiona McDermott. 2016. Articulating co-design in museums: Reflections on two participatory processes. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 13–25.

[17] Steven Conn. 2000. *Museums and American intellectual life, 1876-1926*. University of Chicago Press.

[18] Andy Crabtree, Alan Chamberlain, Rebecca E Grinter, Matt Jones, Tom Rodden, and Yvonne Rogers. 2013. *Introduction to the special issue of "The Turn to The Wild."* ACM New York, NY, USA.

[19] Kirsten M Ellenbogen. 2003. Museums in family life: An ethnographic case study. In *Learning conversations in museums*. Routledge, 92–112.

[20] Exploratorium. *Coding Scheme for APE Multi/Single Station Research Project*. San Francisco, CA.

[21] John H Falk. 2016. *Identity and the museum visitor experience*. Routledge.

[22] John H Falk and Lynn D Dierking. 2000. *Learning from museums: Visitor experiences and the making of meaning*. Altamira Press.

[23] John H Falk, Theano Moussouri, and Douglas Coulson. 1998. The effect of visitors' agendas on museum learning. *Curator: The Museum Journal* 41, 2: 107–120.

[24] Lesley Fosh, Steve Benford, and Boriana Koleva. 2016. Supporting group coherence in a museum visit. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 1–12.

[25] Hugo Fuks, Heloisa Moura, Debora Cardador, Katia Vega, Wallace Ugulino, and Marcos Barbato. 2012. Collaborative museums: an approach to co-design. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 681–684.

[26] Nelson Graburn. 1977. The museum and the visitor experience. *Roundtable reports*: 1–5.

[27] Mark Guzdial. 2013. Exploring Hypotheses about Media Computation. In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research*, 19–26.

[28] Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* 61, 1: 29–48.

[29] Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* 61, 1: 29–48.

[30] Christian Heath and Dirk vom Lehn. 2008. Configuring "Interactivity" Enhancing Engagement in Science Centres and Museums. *Social Studies of Science* 38, 1: 63–91.

[31] Christian Heath and Dirk Vom Lehn. 2004. Configuring Reception: (Dis-) Regarding the 'Spectator'in Museums and Galleries. *Theory, Culture & Society* 21, 6: 43–65.

[32] George E Hein. 2002. *Learning in the Museum*. routledge.

[33] Suzanne Hidi and K Ann Renninger. 2006. The four-phase model of interest development. *Educational psychologist* 41, 2: 111–127. https://doi.org/10.1207/s15326985ep4102_4

[34] Michael S Horn. 2013. The role of cultural forms in tangible interaction design. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*, 117–124.

[35] Michael S. Horn, Erin Treacy Solovey, and Robert JK Jacob. 2008. Tangible programming and informal science learning: making TUIs work for museums. In *Proceedings of the 7th international conference on Interaction design and children*, 194–201. Retrieved October 31, 2015 from http://dl.acm.org/citation.cfm?id=1463756

[36] Eva Hornecker. 2005. A design theme for tangible interaction: embodied facilitation. In *ECSCW 2005*, 23–43.

[37] Eva Hornecker and Jacob Buur. 2006. Getting a grip on tangible interaction: a framework on physical space and social interaction. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 437–446.

[38] Eva Hornecker and Luigina Ciolfi. 2019. Human-computer interactions in museums. *Synthesis Lectures on Human-Centered Informatics* 12, 2: i–171.

[39] Thomas Humphrey, Joshua Gutwill, and The Exploratorium APE Team. 2005. *Fostering Active Prolonged Engagement: The Art of Creating APE Exhibits*. Routledge, Abingdon, UK.

[40] Junko Ichino, Kazuo Isoda, Tetsuya Ueda, and Reimi Satoh. 2016. Effects of the display angle on social behaviors of the people around the display: A field study at a museum. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 26–37.

[41] Natalie Jorion, Jessica Roberts, Alex Bowers, Mike Tissenbaum, Leilah Lyons, Vishesh Kumar, and Matthew Berland. 2020. Uncovering Patterns in Collaborative Interactions via Cluster Analysis of Museum Exhibit Logfiles. *Frontline Learning Research* 8, 6: 77–87.

[42] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK.

[43] Lindsay Lindberg and Ananda Marin. 2020. Designing Dance for Museums: Using Diagrammatic Transcripts to Analyze Embodied Interactions in an Informal Learning Environment.

[44] Duri Long, Takeria Blunt, and Brian Magerko. 2021. Co-Designing AI Literacy Exhibits for Informal Learning Spaces. In *Accepted to Proceedings of The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.

[45] Duri Long, Mikhail Jacob, Nicholas Davis, and Brian Magerko. 2017. Designing for Socially Interactive Systems. In *Proceedings of the 11th Conference on Creativity and Cognition*.

[46] Duri Long, Tom McKlin, Anna Weisling, William Martin, Steven Blough, Katlyn Voravong, and Brian Magerko. 2020. Out of Tune: Discord and Learning in a Music Programming Museum Exhibit. In *Proceedings of the 2020 ACM Conference on Interaction Design and Children (IDC'20)*. https://doi.org/10.1145/3392063.3394430

[47] Duri Long, Tom McKlin, Anna Weisling, William Martin, Hannah Guthrie, and Brian Magerko. 2019. Trajectories of Physical Engagement and Expression in a Co-Creative Museum Installation. In *Proceedings of the 2019 ACM Conference on Creativity and Cognition*. https://doi.org/10.1145/3325480.3325505

[48] Todd Lubart. 2005. How can computers be partners in the creative process: classification and commentary on the special issue. *International Journal of Human-Computer Studies* 63, 4–5: 365–369.

[49] Martin Ludvigsen. 2005. Designing for social use in public places–A conceptual framework of social interaction. *Proceedings of Designing Pleasurable Products and Interfaces, DPPI* 5: 389–408.

[50] Leilah Lyons. 2009. Designing opportunistic user interfaces to support a collaborative museum exhibit. In *Proceedings of the 9th international conference on Computer supported collaborative learning-Volume 1*, 375–384.

[51] Leilah Lyons, Brian Slattery, Priscilla Jimenez, Brenda Lopez, and Tom Moher. 2012. Don't forget about the sweat: effortful embodied interaction in support of learning. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction*, 77–84.

[52] Brian Magerko, Jason Freeman, Tom McKlin, Mike Reilly, Elise Livingston, Scott McCoid, and Andrea Crews-Brown. 2016. EarSketch: A STEAM-Based Approach for Underrepresented Populations in High School Computer Science Education. *ACM Transactions on Computing Education (TOCE)* 16, 4: 14. https://doi.org/10.1145/2886418

[53] Paul Marshall, Yvonne Rogers, and Nadia Pantidi. 2011. Using F-formations to analyse spatial patterns of interaction in physical environments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 445–454.

[54] Kit Martin. 2018. Multitouch NetLogo for Museum Interactive Game. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '18), 5–8. https://doi.org/10.1145/3272973.3272989

[55] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3: 276–282.

[56] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods* 16, 1: 1609406917733847.
[57] Peter Peltonen, Esko Kurvinen, Antti Salovaara, Giulio Jacucci, Tommi Ilmonen, John Evans, Antti Oulasvirta, and Petri Saarikko. 2008. It's Mine, Don't Touch!: interactions at a large multi-touch display in a city centre. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1285–1294.
[58] Jean Piaget. 1955. *The child's construction of reality*. Routledge & Kegan Paul Limited.
[59] David Pinelle and Carl Gutwin. 2008. Evaluating teamwork support in tabletop groupware applications using collaboration usability analysis. *Personal and Ubiquitous Computing* 12, 3: 237–254.
[60] Jessica Roberts, Amartya Banerjee, Annette Hong, Steven McGee, Michael Horn, and Matt Matcuk. 2018. Digital Exhibit Labels in Museums: Promoting Visitor Engagement with Cultural Artifacts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 623.
[61] Jessica Roberts and Leilah Lyons. 2017. The value of learning talk: applying a novel dialogue scoring method to inform interaction design in an open-ended, embodied museum exhibit. *International Journal of Computer-Supported Collaborative Learning* 12, 4: 343–376.
[62] Jessica Roberts and Leilah Lyons. 2017. Scoring Qualitative Informal Learning Dialogue: The SQuILD Method for Measuring Museum Learning Talk. . Philadelphia, PA: International Society of the Learning Sciences.
[63] Yvonne Rogers. 2006. Moving on from Weiser's Vision of Calm Computing: Engaging Ubicomp Experiences. In *International conference on Ubiquitous computing*, 404–421.
[64] Yvonne Rogers. 2011. Interaction design gone wild: striving for wild theory. *Interactions* 18, 4: 58–62.
[65] Jay Rounds. 2006. Doing identity work in museums. *Curator: The Museum Journal* 49, 2: 133–150.
[66] Bertrand Schneider, Patrick Jermann, Guillaume Zufferey, and Pierre Dillenbourg. 2010. Benefits of a tangible interface for collaborative learning and interaction. *IEEE Transactions on Learning Technologies* 4, 3: 222–232.
[67] Phoebe Sengers and Bill Gaver. 2006. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proceedings of the 6th conference on Designing Interactive systems*, 99–108.
[68] Neta Shaby and Dana Vedder-Weiss. Embodied interactions in a science museum. *Science Education*.
[69] Ben Rydal Shapiro, Rogers P Hall, and David A Owens. 2017. Developing & using interaction geography in a museum. *International Journal of Computer-Supported Collaborative Learning* 12, 4: 377–399.
[70] Stephanie Shine and Teresa Y Acosta. 2000. Parent-Child Social Play in a Children's Museum. *Family Relations* 49, 1: 45–52.
[71] Peter K Smith. 1985. The Reliability and Validity of One-zero Sampling: misconceived criticisms and unacknowledged assumptions. *British Educational Research Journal* 11, 3: 215–220.
[72] Barbara J Soren. 2009. Museum experiences that change visitors. *Museum Management and Curatorship* 24, 3: 233–251.
[73] Carey Tisdal. 2004. *Phase 2 Summative Evaluation of Active Prolonged Engagement at the Exploratorium*. Selinda Research Associates, Inc.
[74] Peter Tolmie, Steve Benford, Chris Greenhalgh, Tom Rodden, and Stuart Reeves. 2014. Supporting group interactions in museum visiting. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 1049–1059.
[75] Anthony J Viera, Joanne M Garrett, and others. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med* 37, 5: 360–363.
[76] Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes*. Harvard university press.
[77] Linda L. Werner, Brian Hanks, and Charlie McDowell. 2004. Pair-programming helps female computer science students. *Journal on Educational Resources in Computing (JERIC)* 4, 1: 4. Retrieved April 25, 2014 from http://dl.acm.org/citation.cfm?id=1060075
[78] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology* 13, 1: 61. https://doi.org/10.1186/1471-2288-13-61
[79] Niels Wouters, John Downs, Mitchell Harrop, Travis Cox, Eduardo Oliveira, Sarah Webber, Frank Vetere, and Andrew Vande Moere. 2016. Uncovering the honeypot effect: How audiences engage with public interactive systems. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 5–16.

Received July 2021; revised November 2021; accepted November 2021.