

## The value of learning talk: applying a novel dialogue scoring method to inform interaction design in an open-ended, embodied museum exhibit

Q1 Jessica Roberts<sup>1</sup> · Leilah Lyons<sup>2</sup>

Received: 3 October 2017 / Accepted: 25 October 2017  
 © International Society of the Learning Sciences, Inc. 2017

**Abstract** Museum researchers have long acknowledged the importance of dialogue in informal learning, particularly for open-ended exploratory exhibits. Novel interaction techniques like full-body interaction are appealing for these exploratory exhibits, but designers have not had a metric for determining how their designs are supporting productive learning talk. Moreover, with the incorporation of digital technologies into museums, researchers and designers now have the opportunity for in situ A/B testing of multiple exhibit designs not previously possible with traditionally constructed exhibits, which once installed were difficult and expensive to iterate. Here we present a method called Scoring Qualitative Informal Learning Dialogue (SQuILD) for quantifying idiosyncratic social learning talk, in order to conduct in situ testing of group learning at interactive exhibits. We demonstrate how the method was applied to a  $2 \times 2$  experiment varying the means of control (full-body vs. handheld tablet controller) and the distribution of control (single-user-input vs. multi-user-input) of an interactive data map exhibit. Though pilot testing in the lab predicted that full-body and multi-input designs would best support learning talk, analysis of dialogue from 119 groups' interactions revealed surprising nuances in the affordances of each. Implications for embodied interaction design are discussed.

**Keywords** Dialogue analysis · Museum learning · Embodied interaction design · Exhibit design · Full-body interaction · Intersubjective learning · Human-data interaction · A/B testing

✉ Jessica Roberts  
 jarobert@andrew.cmu.edu

Leilah Lyons  
 llyons@uic.edu

Q2 <sup>1</sup> Carnegie Mellon University, 5000 Forbes Ave., NSH 3602B, Pittsburgh, PA 15213, USA

<sup>2</sup> New York Hall of Science, University of Illinois at Chicago, 1240 W. Harrison St. St. 1535, Chicago, IL 60607, USA

## Introduction

30 Q3

As “designed spaces” for informal learning (National Research Council 2009), museums aim to support learning in a social context, inspire curiosity, and engage visitors with new ideas and phenomena that they can connect to their prior knowledge and understandings. Recent decades have seen a shift in the design of museum environments from a traditional transmission model—in which the museum presents content for visitors to absorb—to a dialogic model (McLean 1999), encouraging visitors themselves to take a more active role in the learning process by incorporating their own values and experiences into their interactions with exhibits (Simon 2010). In addition to using exhibit design to support the more metaphorical “dialogues” visitors have with the exhibits (and by extension, the designers and curators who created them), museums are also more consciously designing exhibits to support the literal dialogues between the visitors themselves (Ash 2004; Atkins et al. 2009; Humphrey, Gutwill, and Exploratorium APE Team 2005; Leinhardt and Knutson 2004).

Full-body, multi-user interactive technologies are of particular interest to museums seeking to engage their visitors in fun, novel activities that support exploratory learning in a social setting. In museums learners can and do bring different objectives for their learning experiences (Rounds 2006), and can even evolve new objectives as their participation unfolds. They work toward these objectives sometimes in concert and sometimes in parallel, but typically with an awareness of each other’s actions. Thus full-body technologies—and the visibility of actions they afford—seem well aligned to support the *social* aspect of a museum interaction, but we do not yet know how well they support *social learning*. In other words: to what extent might full-body interactive experiences facilitate intersubjective meaning-making, per Suthers (2006), wherein “multiple participants contribute to a composition of inter-related interpretations” of a “dynamically changing context”. Much of the educational research into full-body interaction design has explored how bodily movements and perceptions can facilitate an individual’s understanding of abstract concepts, for example, how learners acquire concepts from domains like mathematics (Abrahamson and Sánchez-García 2016; Howison et al. 2011; Davidsen and Ryberg 2017), physics (Enyedy et al. 2012, 2015; Johnson-Glenberg and Megowan-Romanowicz 2017; Lindgren et al. 2016), or even social justice (Antle et al. 2013a), or how movement can help learners interpret abstract representations like graphs (Charoenying 2013; Lyons 2016; Lyons et al. 2012). These designs often focus on how an individual learner’s physical activity can be mapped to an abstract phenomenon or metaphor, an embodied approach where the internal mind-body connection has primacy in the design of the full-body activity.

Some work has demonstrated that full-body interaction can be used in other, more directly social ways, however. Full-body interactives transform human-exhibit interactions into shared performances that can fuel dialogue and reasoning at exhibits (Meisner et al. 2007). In other words, full-body performances can support dialogue by transforming the shared embodied experience into “an object to think with” (Malinverni et al. 2016). While there has been some research into how full-body interactives can facilitate social learning practices like coaching (Tscholl and Lindgren 2016) and other forms of coordination (Malinverni and Burguès 2015), more work needs to be done to better understand the potential benefits of exhibit designs that consciously employ the bodies of visitors as shared, semiotic resources. The role of visitor bodies in meaning-making has been explored by some researchers in art museums, as both an experiential element of appreciating immersive art (Hindmarsh et al. 2005) as

well as a relational tool for engaging with works of art that feature representations of bodies (Steier 2014), but much more can be done, especially in other types of museum settings.

One of the big barriers to fully evaluating the benefits of full-body interfaces for social learning in a museum setting is that interpersonal interactions at exhibits have traditionally been difficult to quantify (Block et al. 2015). Researchers “view one of the richest forms of learning in a museum to be evident in the patterns of discourse and activities that groups engage in - such as labeling, theorizing, predicting, recognizing patterns, testing ideas, and explaining observations” (Atkins et al. 2009), but this very variety and richness poses methodological challenges. Traditional methods for coding and quantifying dialogue, which often structure codes around learning goals, can misrepresent the educational value of an exhibit experience: by focusing only on the museum’s “side” of the visitor-exhibit dialogue, such coding does not capture or honor what visitors bring to the discussion, in the form of their own personal anecdotes or experiences, which we know to be valuable for forming lasting connections to content. Visitors’ spontaneous dialogue during their interactions with an exhibit is presumed to be the primary vehicle for learning (Allen 2002; Ash 2003; Falk and Dierking 2000), yet we have limited means for quantitatively analyzing this dialogue to support in situ A/B testing, which is a critical method for establishing the ecological validity of an exhibit design. In the past, when exhibits were large physically-constructed permanent or semi-permanent structures, such testing was impractical – only one design (that was hypothesized to be best based on formative testing) was built, and all but the most minor changes would be cost-prohibitive after it was installed. Digital exhibits, however, can be recoded for seamless in situ testing of even drastically different designs (Horn et al. 2017), making a new methodology for conducting these tests crucial for conducting research in the modern museum environment.

This paper first describes a methodology, SQuILD, that solves three key problems in quantifying spontaneous visitor talk: identifying socially productive learning talk in open-ended dialogic activities, segmenting spontaneous dialogue to permit cross-group comparisons of talk, and developing a scoring approach that allows alternate styles of learning conversations to be valued. We then describe the application of this methodology to an authentic museum exhibit to reveal how competing interaction designs performed contrary to expectations. While developed to study a museum exhibit, this methodology provides designers with a new technique for studying and comparing CSCL experiences that are intended to have a significant “user contribution” component, wherein the users bring forward and work with ideas of their own, ideas which may not have been specifically anticipated by designers. Techniques for coding and counting dialogue are often used within CSCL because they allow researchers to statistically cross-compare episodes, but the techniques often fail to fully characterize intersubjective meaning-making (Suthers 2006). Intersubjective talk entails learners echoing and reiterating ideas as well as introducing ideas of their own, which depending on the coding categories and counting procedures can either over- or under-represent the learning evidenced by the talk. SQuILD is designed to address these challenges so as to give CSCL researchers a tool that comes closer to characterizing intersubjective meaning-making than existing coding-and-counting techniques.

## Research context

The CoCensus exhibit (Fig. 1) is an interactive census data map display designed to help individuals see a “reflection” of themselves as defined by census data. Census data is a rich

and inherently personal data set—it is, after all, a count of people—yet its complexity renders it nearly inaccessible to casual exploration. Novices require contextualization to make sense of the diverse tallies of the census – numbers alone do not tell a story.

Interactive digital technologies have the potential to introduce visitors to the scientific practice of manipulating large, complex data sets like the U.S. census through visualizations. Researchers are increasingly investigating design challenges of effective data visualizations (e.g. Zhu 2007) including use by non-expert users (e.g. Liccardi et al. 2016) and beginning to attend to how people interpret or misinterpret common data representations (Kay et al. 2016). CoCensus is designed for museums, with the idea that personalizing census data and turning data exploration into a social experience can help spark interest and promote dialogue to engage visitors with what is otherwise a complex and dense corpus of data. Accomplishing this requires presenting the data in a way that is relatable with little or no background knowledge and creating an interactive experience that is as fun as it is purposeful.

### *Fostering relationships with census data*

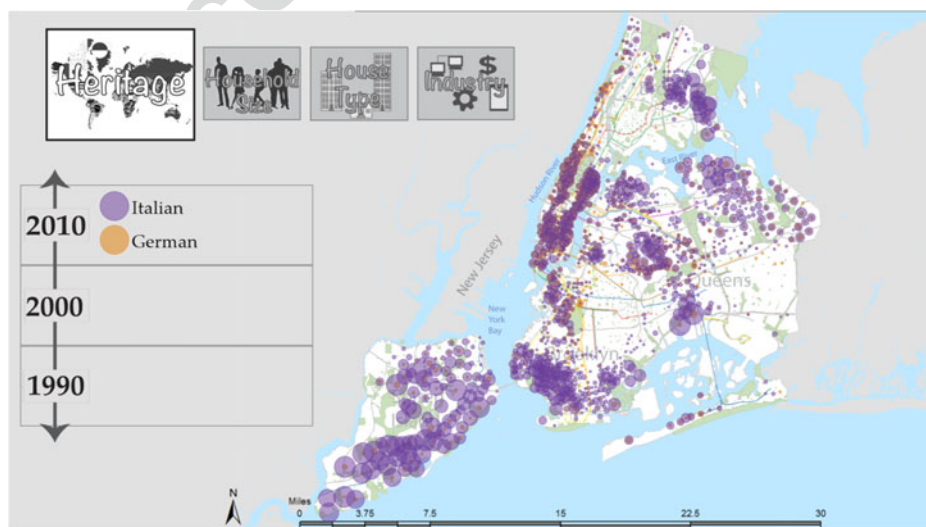
CoCensus displays four categories of U.S. Census data, reflecting different aspects of identity that are relatable to visitors of all ages. Before interacting with the exhibit, visitors are asked to complete a mock “mini-census” survey at a kiosk outside the interaction area (Fig. 2). This survey consists of four questions adapted from the census: 1) What is your ancestry or ethnic heritage? 2) How many people live in your household? 3) In what kind of house do you live? 4) In what industry do you work or want to work?



**Fig. 1** Visitors interact with CoCensus

**Fig. 2** Screen capture of kiosk at which visitors create their census profile. Four census categories were simplified to multiple-choice questions in a graphical user interface

Visitors select answers that represent them or that they find personally interesting, so they are able to explore personally relevant data rather than a museum-curated narrative. When they enter the exhibit interaction area, each visitor will see scaled centroids (“bubbles”) on a large map display representing the numbers of people who answered the question the same way as that visitor (e.g., people who work in the Transportation industry). Spatializing data by overlaying it onto a map of the local geography affords visitors the opportunity to relate the data to authentic lived experiences of themselves and their companions. Two visitors interacting together can compare the size and distributions of their populations (each represented by the visitors’ self-selected bubble color, see Fig. 3.).



**Fig. 3** The CoCensus exhibit displays multiple visitors’ self-selected data as scaled centroids (“bubbles”) overlaid onto the local geography



## Physically engaging visitors through interactivity

150

The conceptual connections fostered by customization and spatialization of the data are augmented with a physical connection – providing users control over the representation. Allowing learners to manipulate visualizations can provide unique affordances for recognizing patterns (Card et al. 1999), and giving agency to users lets them explore what they find most interesting. Through a multi-year design-based research program we have investigated multiple aspects of the interactivity (Cafaro et al. 2014b, 2013; Roberts et al. 2014). The exhibit iteration discussed here affords three manipulations to the visualization: selecting a category of census data (heritage, household size, housing type, or industry); choosing a census year (1990, 2000, or 2010); and setting the aggregation level of the data (census tract, borough, or city-wide, see Fig. 4).

It was an open question how best to allow visitors to perform these manipulations. As introduced above and discussed in more detail below, embodied controls—including full-body interactivity—seemed appealing for the potential to support intersubjective learning. Early testing of full-body interaction designs showed great promise in a lab setting (Cafaro et al. 2013) and in situ (Roberts et al. 2014), but we couldn't yet verify whether full-body interactivity would support more variety and depth of learning talk than other means of controlling the system, such as a handheld tablet. We also didn't know whether individualized control (each user independently manipulating her own representation) would produce greater learning outcomes than a global controller allowing a single user to manipulate all data bubbles together. These two factors, the *means of control* (full-body or handheld) and *distribution of control* (single or multi input) have significant implications for the final implementation of the exhibit, so we needed to experimentally determine the most favorable interaction design. Standard metrics used in A/B testing of digital interfaces, such as hold time and click counts, don't speak to how well the designs are accomplishing the main objective of this exhibit – getting people talking to each other about data. How could we compare dialogue by visitors of varying ages, interests and background knowledge who saw unique and customized datasets, the representations of which they selected themselves? We needed a new methodology for understanding this idiosyncratic talk, where the inter-related interpretations *constitutes* the learning (Suthers 2006).

## Background

180

Sociocultural theories of learning posit that learning is a product not only of individual cognitive processes but of the broader cultural, historical, and institutional context, asserting that “analytic efforts that seek to account for human action by focusing on the individual agent

181

182

183



**Fig. 4** The CoCensus display at three aggregation levels of data: census tract (left), borough (center), and city (right). Each representation affords different kinds of spatial and quantitative comparisons between the datasets

are severely limited” (Wertsch 1998). This perspective follows the Vygotskian view that humans virtually always learn through dialogues, asking questions, and negotiating meaning (Vygotsky 1978). Our study follows a large body of work adopting a sociocultural perspective for museum learning (Allen 2002; Ash 2003; Atkins et al. 2009; Crowley and Jacobs 2002; Leinhardt et al. 2002; Schauble et al. 2002; Steier et al. 2015) and views visitors’ meaning making as a form of mediated action. The agents (the museum visitors) are constructing meaning not individually, but through the interactions between themselves and their mediational means: other agents in the space, the exhibit and its control mechanisms, and the perspectives they employ during the interaction. It is through the interactions between individual and shared understandings through dialogue that meaning is collaboratively made. Therefore the measure of learning considered by this research is visitors’ “learning talk” (Allen 2002) produced during their interaction.

### Supporting learning dialogues in museums with exhibit design

As the museum exhibit design community began embracing more social exhibit designs that were intended to support visitor-visitor dialogue in addition to visitor-content dialogic interactions (Ash 2004; Atkins et al. 2009; Humphrey et al. 2005; Leinhardt and Knutson 2004), a number of exhibit designers began to turn to technology to facilitate these (often simultaneous) dialogues. And as technologies have become increasingly ubiquitous, there has been an explosion in the variety of designs to support social exhibit experiences, using technologies as varied as large shared displays (e.g., Diamond et al. 1995; Macedonia 2003; Meisner et al. 2007), handheld mobile devices (e.g., Aoki et al. 2002; Cabrera et al. 2005; Hope et al. 2009; Lanir et al. 2016; Yatani et al. 2004), combinations of shared displays and mobile devices (e.g., Dini et al. 2007; Kruppa and Aslan 2005; Lyons 2009; Scheible and Ojala 2005), interactive tabletops (e.g., Angelo et al. 2015; Antle et al. 2013b; Davis et al. 2015; Lyons et al. 2015) augmented reality (e.g., Beheshti et al. 2015; Schmitt et al. 2010; Wagner et al. 2006; Yoon et al. 2012), and full-body interaction (e.g., Roberts et al. 2014; Tscholl and Lindgren 2016; Uzzo et al. 2016; Yap et al. 2015). The foregoing designs were all intended to afford different social behaviors – from spectating (e.g., Diamond et al. 1995; Tscholl and Lindgren 2016) to coordinating (e.g., Lyons 2009; Angelo et al. 2015) to co-constructing (e.g., Antle et al. 2013b; Lyons et al. 2015) – but vary in the degree to which the researchers examined the implications of their designs for learning behaviors and outcomes. Full-body interaction exhibits are becoming increasingly popular in museums, both because in some ways they can be easier to deploy and maintain (fewer mechanical parts that visitors can scratch or damage, no charging or other management of mobile devices needed) and because they may more naturally support simultaneous use by groups of visitors. There is a need for more CSCL research to explore how full body designs can support learning as well as social interactions.

### *Full-body interaction and situated learning*

Full-body interaction is often described as a form of “embodied interaction” following Dourish’s definition, which calls for designers to attend to the full scope of how a user is situated within a context: within the user’s body’s own sense of itself and its motion (proprioception/kinesthesia), within the physical setting, and within the social setting (Dourish 2001). As a consequence, it can be confusing when confronted with competing claims for how the “embodied interaction” found in full-body interaction designs can support

learning, because the learning supported by two different designs might be relying on different mechanisms. To clarify, we found it helpful to borrow from the Robbins and Aydede (2009) description of the three major theses found in situated cognition: (1) *embodiment*, meaning that cognition is rooted not just in the brain but in the body and its actions and sensations; (2) *embedding*, meaning that cognition exploits structures found in the natural and social environment; and (3) *extension*, which examines how cognition can span beyond individual organisms. So, designs that rely on gestalt perception (Gibson 1979) and affordances (Norman 1988) hew more closely to the *embodiment* thesis; while designs that attend to how a learner and his or her actions are situated (Suchman 1987) and mediated (Wertsch 1994) by physical or sociocultural aspects of the setting are more strongly aligned with the *embedded* thesis; while designs that promote learning activities that are distributed across multiple participants and tools (Engeström et al. 1999; Seifert and Hutchins 1992; Vygotsky 1978) adhere more closely to the *extension* thesis. Using this framing, one can see how designs often lumped together under the heading of “embodied interaction” pull more strongly from one or more of these three theses. The next three sections will review existing “embodied” educational designs using each of these theses in turn.

### **Embodied full-body designs for education**

The majority of research into the educational value of full-body interaction has taken more of an *embodied* approach, where the internal mind-body connection has primacy in the design of the full-body activity. For example, researchers have investigated how learners acquire mathematics concepts by positioning their bodies to enact abstract concepts like ratios (Howison et al. 2011) or scale (Fischer et al. 2011), thus constructing an abstract understanding from bottom-up sensorimotor schema (Abrahamson and Sánchez-García 2016). Researchers creating full-body interactives for physics education often use a similar *embodied* approach, asking learners to enact phenomena like trajectories of meteors (Tscholl and Lindgren 2016) or balls (Enyedy et al. 2015) by running along a path, or to trace particle vectors with gesture (Johnson-Glenberg and Megowan-Romanowicz 2017). In all of these cases, motor engagement is critical to the learning – the act of movement is a direct enactment of the conceptual schema.

The designs above rely on a tight congruency between the full-body movement and the abstract concept (Johnson-Glenberg et al. 2014), but there are examples of looser congruency being used to help participants learn and employ abstract metaphors in areas as diverse as social justice (Antle et al. 2013a), music (Antle et al. 2008), and comprehension of argumentation in text (Kaschak et al. 2017). With all of these “enactive metaphors” (Gallagher and Lindgren 2015) learners acquire an analogic understanding of a concept via enacting it, similar to “image schema” (Lakoff and Johnson 1980).

Image schema (metaphors like “up is more” or “back in time”) are the embodied metaphors people form as a result of their earliest sensorimotor interactions with the world, and are thought to be present in most humans thanks to our shared physiology (Lakoff and Johnson 1980). Thus, rather than using full-body interaction design to help learners build a new concept or a metaphor, some *embodied* design researchers have approached the problem from the opposite direction by exploring if these pre-existing image schema metaphors can be used to engineer full-body interactions that are more usable (Antle et al. 2009) or that can encourage learners to assume different perspectives (Roberts et al. 2014).



Finally, some researchers have similarly used *embodied* design strategies to explore how to help learners interpret abstract representations, linking differences in hopping frequency (Charoenying 2013) or playground step counts (Lee and Drake 2013) to graphs of activity, or perceived physical exertion to graphs representing the variability, magnitude, and rate of climate change (Lyons 2016; Lyons et al. 2012). In these examples, the congruency is between the physical action and aspects of the visual representation – while the visual representation does not act as an “attentional anchor” as it does in highly coordinative full-body enactments (Abrahamson and Bakker 2016), it does act as an intermediary, helping learners connect their physical movements or proprioceptive sensations to elements of abstract representations like outliers and variability.

### ***Embedded full-body designs for education***

The preceding *embodied* full-body designs may have given primacy to the mind-body connection in structuring the full-body learning activity, but several of them also attended to how full-body movements are *embedded* within an external physical and social context. For example, visual graphical representations are necessary to support learners’ ability to perceive outliers and variability in performance data (Lee and Drake 2013; Lyons 2016), because these phenomena arise from the aggregation of a number of learners’ enactments. A learner’s singular physical experience of their own performance might give a window into the representational space, but does not map onto the totality of the represented phenomenon – the performance must be *embedded* in physical representation to allow learners to detect and reflect upon the targeted concepts. In addition to supporting reflection, some physical design decisions affect how learners direct future activities. For example, the presence of projected representations of meteor trajectories on the floor (Lindgren et al. 2016) or ball forces superimposed on video feeds (Enyedy et al. 2012) assist learners in interpreting their actions and the effects of their actions. This support provided via the visualizations is a form of cognitive offloading that has been conceptualized as a “liminal blend” (Enyedy et al. 2015), wherein the sensorimotor/conceptual mapping targeted by the *embodied* design is enhanced by an “overlay” of visualizations and physical objects that learners can use to develop their conceptual understanding, allowing them to tune their performances to come closer to enacting a targeted phenomenon.

Social embedding does not require that a full-body interaction design re-present the performance of the enacting learner – the performances themselves (i.e., how the physical actions are structured) can serve to embed the learner more thoroughly within a social context. Research has shown that the design of exhibits can help transform human-exhibit interactions into shared performances (Meisner et al. 2007), which can offer a number of potential benefits for the social learning experience. For example, studies have shown how spectating visitors gain a valuable “preview” of how to interact with an exhibit by watching current exhibit users (vom Lehn et al. 2001), and that witnessing companions’ actions at a shared exhibit can spark conversations that support learning (Lyons et al. 2015). At a very simplistic level, full-body interaction does not get “in the way” of usual human communication. A number of museum professionals are concerned by the tendency of digital technologies (sometimes pejoratively called “screens”) to capture visitor attention to the extent that visitors no longer attend to and interact with companions (Lyons 2009). Full-body interaction is seen as a way to help visitors keep their “heads up” and thus preserve visitor-visitor dialogue by remaining embedded in the social context. In full-body interaction designs where the emphasis is on *embedding* the learner

in a social setting, the *embodied* notion of tying an enacting learner's sensorimotor performance to a concept might be loosened or disregarded entirely. Rather, designing a full-body interaction that supports communication is key.

### ***Extended full-body designs for education***

There is great potential for going beyond preserving to enhancing visitor-visitor dialogue through creative full-body interaction design. For example, visualizations that *embed* a full-body performance within a physical context can also be used to *extend* the cognitive activity to include others in addition to the enacting learner, creating the possibility to support distributed cognition or scaffolding. Tutors (Abrahamson and Bakker 2016), classmates (Enyedy et al. 2015), and parents (Tscholl and Lindgren 2016) have successfully used visualizations of embodied performances to help coach the learners. Interpretive staff at a zoo used visualizations of visitors' full-body performances as a jumping-off point to help both the enacting learner and the watching audience to relate the performance to a range of concepts, from the lived experience of polar bears in the wild to the mathematical concepts underpinning climate change (Slattery et al. 2014). Even in the absence of visualizations of enactments, full-body interaction translates the locus of visitor interaction with digital exhibits from "screen space" into the physical, social space (Cafaro et al. 2014a), making visitor actions and choices more visible. Some full-body interaction designs take the form of participatory simulations, engaging learners as bees (Peppler et al. 2010) or residents of an ecosystem (Uzzo et al. 2016) to co-construct an emergent system via their actions. By intentionally employing the bodies of visitors as shared, semiotic resources that visitors can reason with and around, a full-body interaction experience can be *extended* into "an object to think with" (Malinverni et al. 2016). *Extended* full-body interaction designs also *embed* learners in a social context, but some of the normal body positioning, pointing gestures, and symbolic gestures humans use as semiotic resources within their conversations with others (Goodwin 2000) might be sacrificed in favor of encouraging learners to enact physical actions that are highly recognizable and visible (but might be mildly artificial). We took this latter approach in the full-body version of the CoCensus exhibit: visitors enact symbolic gestures to control the exhibit.

### ***Situating full-body interaction research in museums***

The different situated cognition mechanisms (embodied, embedded, or extended) that can be at play in full-body interaction designs have major implications for how to study the learning in those experiences. Full-body interaction researchers have recommended first studying an interaction design in the lab, only moving the research to the in situ location once the operative mechanisms have been identified (Lindgren and Johnson-Glenberg 2013). While this might work well for full-body interaction designs with a strong *embodied* focus, designs that rely more heavily on the *embedded* or *extended* aspects of situated cognition are much harder to examine in this way, because the physical setting, social setting, and interactions between learners are essentially co-constructing the learning experience. Highly interactive museum learning experiences are notoriously hard to study outside of their intended context of use – in practice, the different design features can interact in unpredictable (and often undesirable) ways (Allen 2004). To move forward with productive interaction designs for engaging visitors, then, informal learning researchers and designers need to understand how different mediational means work together—or in conflict—to affect learning outcomes. Researchers need to be able

to conduct experiments that vary the combinations of different design features, and to conduct those experiments, they need to develop methods to document and measure the emergent learning that occurs at open-ended museum exhibits.

## Measuring informal learning talk: The SQuILD method

Many scholars have addressed the challenge of measuring and quantifying learning in museums. Some efforts have tried to measure individual knowledge gains via pre- and post-tests or interviews (e.g., Falk and Storksdieck 2005). The accuracy of such measures depends on visitors accessing that content during their visit, which is not assured to happen where visitors pick and choose paths through an exhibition (Diamond et al. 2009). It is even less reliable in customizable exhibits like CoCensus where visitors control what they see and how they see it – even if questions were tailored to content they had seen. Exit interviews are also problematic, as “there is little correspondence between people’s post hoc characterizations of their experience and the activities in which they engage when visiting exhibitions” (Heath and vom Lehn 2008). Some research efforts have explored how observational data can serve as proxies for learning. For instance, Borun et al. (1996) linked videorecorded behaviors of family groups on a museum visit with the groups’ exit interviews about the exhibit content, and found the frequency of certain observable “significant behaviors” could distinguish which families had richer content engagement. Such measures and other countable metrics like hold time have been used to experimentally test and compare multiple versions of an exhibit, for example with and without augmented reality components (Yoon and Wang 2014).

From a sociocultural perspective, however, the most reliable and accurate indicator of learning during an exhibit experience lies in the dialogue among group members as they interpret content together. Schemes by researchers like Allen (2002) and Atkins et al. (2009) have defined important conversational elements in museum “learning talk” (Allen 2002), but they tell us only what to code, not how to create a measurement that reliably permits statistical comparisons across conditions. We have identified three primary challenges associated with identifying *and* measuring productive talk. Here we outline those challenges and describe a novel methodology, SQuILD (Scoring Qualitative Informal Learning Dialogue), for addressing them.

### Challenge 1: Identifying open-ended learning with interactive exhibits

Researchers have identified a number of features of productive learning talk, for example reading text aloud, asking and answering questions, connecting new information to prior knowledge, and giving explanations to companions (Allen 2002; Ash 2003; Leinhardt and Knutson 2004; Kisiel et al. 2012). This prior work laid an important foundation, but these schemes were largely developed before exhibits were highly-interactive. To study open-exploration interactive exhibits, these schemes should be consolidated and extended to embrace visitors’ agency in shaping their own interaction experiences.

### Addressing challenge 1: Identifying learning in open-ended informal dialogue

We began by consolidating common threads in the literature to develop five categories of substantive visitor talk applicable to interactive exhibits—management, instantiations,

evaluations, integrations, and generations—described further below. We then used an open coding process, informed by the literature of our exhibit’s content domain (data interpretation), to determine specific sub-codes within each of those categories (detailed in Table 1 below). While our sub-codes were specific to our exhibit content, the five categories could be adapted to a variety of interactive exhibits. Future work embracing this methodology would likely find that developing unique sub-codes within these presented categories would retain the structure of the method while adapting to the specific content focus of the research.

Manage codes

When multiple people are interacting with an exhibit, some talk will directly address the interaction with the exhibit. Talk related to the establishment of joint attention, negotiation of action, or scaffolding exhibit use was coded as *management*. These kinds of behaviors are of interest because they speak to how visitors are working together and mediating each others’ experiences. For example, Allen (2002) categorized these kinds of actions as “strategic” with only two sub-codes: “use” and “metaperformance.” Borun et al. (1996) attended to observable coordination behaviors like “call over.” Multiple studies have attended to facilitative behaviors such as explaining, asking and answering questions, and suggesting actions (Ash 2003; Eberbach and Crowley 2005; Diamond et al. 2009; Atkins et al. 2009), all of which fall under *management* in our scheme. Researchers of technology-based multi-user interactives who are concerned with interpersonal interactions like interference (Falcão and Price 2009), negotiation of exploration (Davis et al. 2015), and joint discovery of interactive possibilities (Williams et al. 2005), could detect and characterize these behaviors via patterns in *management* codes.

Instantiate codes

The term “instantiation” indicates when a user says aloud a piece of information, providing opportunities for other visitors to internalize that information (i.e., learn from the exhibit). Per sociocultural learning theory, learners must articulate ideas via communication before individual learning can take place through internalization or personalization (Vygotsky 1978).

**Table 1** Sub-codes were sorted according to their relevance to learning objectives

Low Relevance (1)	INSTANTIATE category INSTANTIATE dataset INSTANTIATE decade INSTANTIATE geography INSTANTIATE representation INSTANTIATE self	INTEGRATE connect multiple INTEGRATE connect simple MANAGE ask interpretive question MANAGE direct co-visitor’s movements	MANAGE narrate intentionality MANAGE negotiation of control MANAGE purpose of exhibit MANAGE suggest action
Mid Relevance (2)	EVALUATE characterize EVALUATE win GENERATE contextualize GENERATE identify knowledge gap	INTEGRATE challenge interpretation INSTANTIATE outside knowledge	MANAGE ask guiding question MANAGE clarify MANAGE direct co-visitor’s attention
High Relevance (3)	EVALUATE question census categories GENERATE confirm GENERATE make prediction	GENERATE negotiate meaning GENERATE notice surprising pattern	GENERATE pose inference INTEGRATE compare

Instantiations are an important part of the social learning process as they can lay the foundation for further reasoning among learners on a museum visit (Kisiel et al. 2012) and help establish joint attention (also referred to as “grounding”). Processes of noticing and establishing joint attention among visitor group members have been found to be productive in facilitating learning talk in museums (Povis and Crowley 2015; Leinhardt and Crowley 1998), and reading labels aloud was identified as a “significant behavior” linked to increased group learning by Borun et al. (1996).

## *Evaluate codes*

Evaluation statements make a judgment or assessment about a piece of information by assigning some kind of value, whether qualitative or quantitative. Such personal qualitative evaluations are very important in informal learning settings, where developing one’s identity is seen as just as much of a goal of the meaning-making process as absorbing content (Rounds 2006). In this context, evaluations can be simple standalone comments (e.g. “They are all spread out!”) or part of a more complex statement. The most common *evaluate* sub-code in our exhibit was *characterize*. *EVALUATE-characterize* statements could be quantitative (e.g., “a lot” or “not very many” of something), or spatial, e.g. describing a population as being “everywhere.”

## *Integrate codes*

While evaluation statements refer to a single idea, the final two categories connect multiple pieces of information in some way. Friel et al. (2001) refer to the act of looking for relationships in data as “interpretation.” The SQiILD framework adopts the more precise term *integration* from Murray et al. (1998) to describe the act of pulling together multiple pieces of information presented in an exhibit. Statements that integrate are those that make explicit connections or comparisons between multiple pieces of information: for example, in our exhibit, between two different datasets, between a dataset and the geography, between a dataset and itself over time, etc. Connections and comparisons are integrative talk widely acknowledged to be valuable in museum settings (e.g. Allen 2002; Atkins et al. 2009; Falk and Dierking 2000.)

## *Generate codes*

*Generate* statements “[go] beyond the data” (Curcio 1987) to combine information from the exhibit with visitors’ own prior knowledge and experiences. Falk and Dierking’s (2000) Contextual Model of Learning posits that what learners gain during a learning experience is inextricably tied to the personal context they brought into the experience—prior knowledge, experiences, motivations, identities, etc. Allen (2002) incorporates what she calls “connecting talk” into her framework for analyzing visitor conversations at an exhibit, but unlike the *connections* described above as an *integrate* code, the type of connections she is referencing are making use of outside information, by connecting an exhibit to life, prior knowledge, or other exhibits. She describes this stitching together of information from different sources as “powerful and ubiquitous means of learning in informal settings.”



## Summary

466

Taken together, these five categories of codes capture the substantive talk visitors have during exhibit interactions. These categories consolidate prior work studying museum dialogue and are likely to apply to any interactive exhibit. The subcodes (Table 1) are exhibit-specific, so researchers employing this method will likely adapt these and create new codes according to their unique content and context. The next section discusses how these codes can be applied to visitor dialogue.

467  
468  
469  
470  
471  
472

## Challenge 2: Segmenting dialogue to permit cross-group comparisons

473

The grain size for segmentation is a key decision in any analytical process. Due to the spontaneous nature of joint exploration, many ideas are split among two or more visitors as they work together to make sense of the content. Visitors interrupt each other and in some cases interrupt themselves mid-idea as they notice new information. The fragmented nature of museum dialogue is a known challenge for assessing learning in this context (Allen 2002), particularly when an analysis aims to quantify talk by counting instances of a particular kind of speech act. Some analyses of visitor dialogue address this challenge by coding simply for the presence or absence of a particular kind of talk (e.g., making a prediction) at all during a session (Allen 2002; Atkins et al. 2009), but such an analysis runs the risk of skewing the quantification toward under-representation: a visitor group that had an in-depth conversation with many predictions would receive the same score as a group that made only a single prediction. Another common segmentation strategy, one that supports valid comparisons of sessions, is to divide speech into conversational turns and code and count those turns to quantify them (Chi 1997). Because of the frequent interruptions and repetitions common in informal learning talk, this delimitation technique could skew the quantification toward over-representation of certain kinds of talk. Larger delimitations, meanwhile, such as segmenting by theme or referenced data, would obscure the intricacies of the productive dialogue. A new segmentation method is necessary.

474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493

## Addressing challenge 2: Segmenting dialogue through idea units

494

Dialogue is a group activity. Some ideas are spoken by only one visitor and are contiguous and completed in a single conversational turn. Others are co-constructed by multiple visitors as they collaboratively investigate the exhibit's content. To reach the appropriate level of granularity, this method adopts the *idea unit (IU)* as its unit of analysis, introduced by Jacobs et al. (1997) as "marked by a distinct shift in focus or change in topic." We amend this to more closely capture dialogue emerging in the midst of a group activity by defining an idea unit as marked by a distinct shift in focus or change in topic *or purpose*. This adjustment segments visitor conversation into chunks according to what that speech is doing in the group interaction. Idea units can range in length from a single word, e.g., reading aloud a category name, to a multi-sentence utterance. To illustrate the concept, below are two excerpts of dialogue from

495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505

two visitor sessions. The first shows somewhat straightforward linear idea units, as annotated below:

- |        |  |   |
|--------|--|---|
| [1] A: | I want to see how it changes.  | [states intention]                                      |
| [2] A: | Like that area over there changed a lot in regards to... demographics, you see it? | [draws joint attention to areas that changed over time] |
| [3] B: | Yeah.  |   |
| [4] A: | And up there.  |   |
| [5] B: | More spread out.   | [characterizes data]                                    |
| [6] A: | But you see the greatest change here on this side.                                 | [identifies area of particular interest]                |

This excerpt was divided into four idea units. Lines 2–4 were considered to be one idea unit because the fast pace of these conversational turns indicated that to the participants they contributed one chunk of information. The same words spoken more slowly with significant pauses in between turns could have been separated into individual units. This reliance on pacing in the moment—how each speech unit would have been perceived by users in the experience—illustrates the importance of completing this segmentation from videos rather than from transcripts, as discussed in more detail below. The idea units in this example vary in length and in one case span multiple turns and speakers, but they are fairly straightforward. Some idea units are less obvious, because they are detached and inter-spliced. Take this segment from another pair:

- A: So whatever's, I'm assuming there must be railway or, oh wait, isn't that a road? That goes across, across the water. So there's-
- B: It's a bridge.
- A: So my guess is, oh it's a waterway or a roadway or whatever. Waterway maybe. But that area's most likely industry.

Visitor A's main goal is to pose his theory about the area being industrial but he keeps interrupting himself trying to correctly describe the roadway, with an interjection from his companion. This segment is counted as two overlapping idea units, as the participants are doing two meaning-making moves in these three turns: decoding the map representation, represented with an underline below, and posing an inference about the area based on the data ("So whatever's... So there's... So my guess is... But that area's most likely industry," double-underlined below).

- A: So whatever's, I'm assuming there must be railway or, oh wait, isn't that a road? That goes across, across the water. So there's-
- B: It's a bridge.
- A: So my guess is, oh it's a waterway or a roadway or whatever. Waterway maybe. But that area's most likely industry.

This segmentation into idea units prevents stutters and echoing (e.g., the repeated starts to the inference "So whatever's", "So my guess is...") from unfairly weighting a statement beyond its contribution to the dialogue, which can occur in a speaking-turn-based quantification of talk (Chi 1997). Idea-unit coding is particularly useful when characterizing the overall educational quality of a group's conversation, rather than trying to draw attention to the individual contributions or cognitive acts of each speaker. Given the sociocultural perspective

much work in museum learning is taking (namely, learning is evidenced in the group's talk, and benefits the group as a whole), idea units are more appropriate than a turn-based approach.

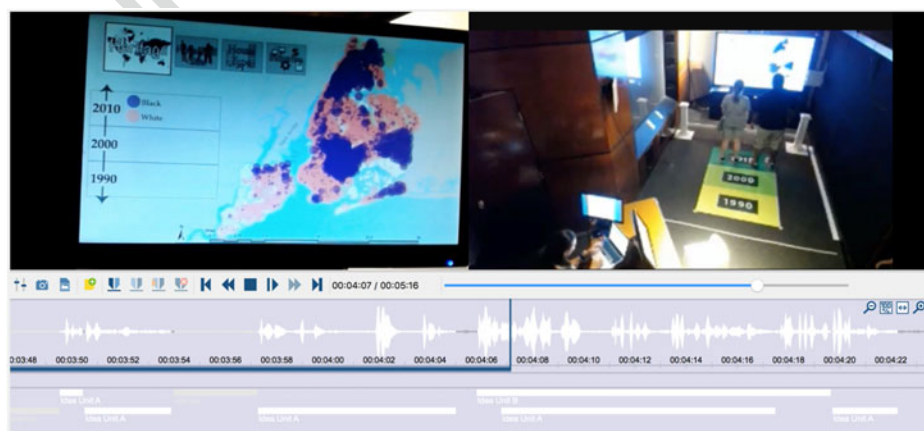
Idea-unit coding is best done directly from video to retain the context of visitors' comments (see Fig. 5). Separating dialogue from the visitors' experience by making—and later coding from—a transcript removes the context in a way that obfuscates or even completely alters the meaning of the statement. Context is particularly important for learning talk occurring at dynamic interactive exhibits, where visitors can alter the exhibit state with their actions, and their dialogue responds to the changing state of the display.

### Challenge 3: Respecting socially-constructed learning when quantifying dialogue

In addition to the coding and segmentation challenges described above, a further complication is the irreducible tension that while not all dialogue may be equally well aligned with exhibit learning goals, all visitor groups are guaranteed to engage with exhibits in manners that suit their current interests and level of understanding. What constitutes meaningful learning talk for one group might not align with content goals for the exhibit. For example, it is widely accepted that reading a label aloud is a productive form of talk in museums (Borun et al. 1996; Kisiel et al. 2012; Allen 2002; Atkins et al. 2009), but should such reading merit the same quantitative score as a comparison of two datasets, if the exhibit's aim is to foster such comparisons? But conversely, should a "conversation" where one visitor makes a single comparison be valued more highly than an extended dyad conversation where both make observations about exhibit content? An analysis should ideally respect the socially constructed nature of museum learning and "give credit" to both quality and quantity of talk, acknowledging all productive talk while retaining qualitative distinctions among different kinds of talk.

### Addressing challenge 3: Quantifying depth and nuance in visitor talk

Quantifying visitor talk begins by assigning the five categories of codes identified above to the identified idea units. Any idea units that did not fit into a category and sub-code were marked *non-substantive* and were disregarded in the analysis. Some idea units were coded with a



**Fig. 5** Coding directly from video maintains the context of the talk and facilitates segmentation of overlapping idea units, shown here as white bars in the coding software MaxQDA

single code. Many idea units, however, were coded with multiple codes: though the statements were one logical idea, they were deep and complex enough to warrant multiple codes. This process of simultaneous coding (Saldaña 2009) maintains the richness of the talk, rather than reducing a statement to a single code. For example, here one participant compares two groups from the Heritage category, those who had identified as White, and those who had identified as Mexican<sup>1</sup>:

“Okay, there’s a lot of White in 2000, rather than Mexicans.”

This statement as an overall idea unit compares the two heritage groups. Within that broad goal, it does multiple things. It *INSTANTIATES* the *datasets* (“White” and “Mexicans”) and *decade* (“2000”), it *INTEGRATE-connects* the dataset to the decade “White in 2000”, it *INTEGRATE-compares* the datasets (“White” and “Mexican”) using “rather than,” and it *EVALUATE-characterizes* “White” as being “a lot.” Coding the statement only as a single code—in this case *INTEGRATE-compare*—would give it the same value as a much less rich statement like, “It looks like there are more of them.” Only via simultaneous coding can we give credit to the multiple “hooks” this complex statement provides for further discussion.

## *Assigning values to codes by tying them to exhibit learning goals*

The coding framework described above flags conversational acts that are likely to contribute to shared meaning making at an interactive data exhibit. In any open-ended exhibit, multiple kinds of talk are considered to be highly relevant to the intended learning. Other kinds of talk are important but less directly aligned with the learning goals. Therefore, this methodology employs a form of magnitude coding (Saldaña 2009; Miles and Huberman 1994) by sorting the sub-codes into high, medium, or low categories according to their relevance to the goals of the exhibit, as determined by the research team (see Table 1; note that the weighting is particular to this exhibit). Using magnitude coding as a way of “quantizing” a phenomenon (Tashakkori and Teddlie 2010) permits the use of inferential statistics (Bernard 2006; Saldaña 2009) in order to compare the experimental conditions.

The subcodes assigned to the “low relevance” category—mostly *INSTANTIATE* and *MANAGE*—are all activities that are useful for grounding and coordinating the group learning experience and may serve as springboards for future dialogue, but are themselves not strongly related to the learning goals of the exhibit. These statements were assigned a weight value of one. “Mid relevance” codes took steps to make sense of the presented data by characterizing and contextualizing it, clarifying the representational forms, and directing co-visitor’s attention to an interesting element of the exhibit (which rises above a simple *stantiate* code because it conveys to the listener that the targeted element is worthy of joint discussion). These codes were given a weight score of two. “High relevance” talk included statements that related presented data to prior knowledge or expectations, predicted or inferred information, compared

<sup>1</sup> The complex and often contested way the U.S. Census counts heritage information has posed numerous design challenges in creating this exhibit that have been discussed elsewhere (Roberts et al. 2015). In the iteration of the exhibit tested here, the designations provided by the census are preserved. In the Heritage category, visitors had the option of selecting one option from any of the following categories: Race (e.g. “White” or “Japanese”), Hispanic status (e.g. “Mexican” or “Puerto Rican”), or Ancestry (e.g. “Arab” or “German”). This often resulted in pairs of users exploring mismatched datasets when one user chose a race and another an ancestry or Hispanic group.

datasets with each other or over time, and questioned the source of the data (such as how the census counts a particular category). This kind of talk is exactly the kind of exploration and meaning making the exhibit is intended to support, and thus were assigned a weight of three. Overall “content scores” were calculated by summing the weighted values of all codes applied to a session.

### *Validity and limitations of magnitude coding*

The assumption being made by this approach is that the overall richness of codes corresponds to the overall richness of the shared learning experience throughout the session, and that idea units are used to avoid over- or under-representing that richness. There is no assumption that a session’s value should be determined by the tally of “high value” (i.e. multi-coded) idea units like the example above, or that calculating the average value of idea units over a session is a meaningful measure. Other methodologies exist to closely scrutinize individual discourse statements. Instead, we recommend summing all codings applied over a session to assign a quantified value to what learners were able to do in the session. Looking at another example:

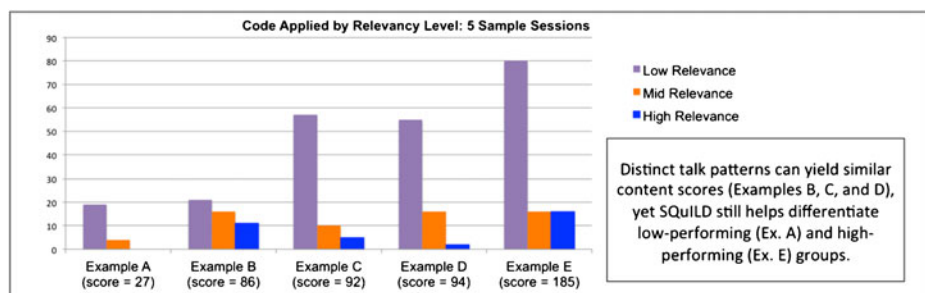
“Oh yes, lots of West Indians in Brooklyn, that is true.”

This statement *INSTANTIATES-dataset* (“West Indians”, weight of 1) + *INSTANTIATES-geography* (“Brooklyn”, 1) + *EVALUATES-characterize* (“lots of”, 2) + *INTEGRATES-connect:simple (dataset to geography)* (West Indians in Brooklyn”, 1) + *GENERATES-confirm* (“oh yes ... that is true”, 3) = content score of 8. Whether delimited as one 8-point idea unit versus a 3-point idea unit (“Oh yes... that is true”) plus a 5-point idea unit (“lots of West Indians in Brooklyn”), the impact on the session score is the same. Because of this flexibility in segmenting idea units, the methodology does not recommend analyzing scores of individual idea units (e.g., to compute metrics like “average value per idea unit”), but only the total dialogue in a session.

Session scores illuminate differences between conditions in their ability to support visitors in productive exploratory talk. Even the codes identified as “low relevance” are still productive learning talk. In this context, high-value codes often (but not always) build on low and mid-value talk, and a good statement often contains all three. A simple IU like “It looks like there are more of them” is a high-level *INTEGRATE-compare*. But it adds far less substance to the conversation than a richer statement like the 8-point IU described above. The simpler high-level statement, which our method would assign a value of 3, gives the speaker’s companions fewer “hooks” to build on: they can only respond to the comparison, whereas the more complex statement gives companions a number of different directions to take the conversation. Given the open-ended nature of the interactions and the underlying assumption that each group will be approaching the exhibit from a unique background and with unique goals, it is to be expected that productive interactions will not be the same for each group. Five example sessions graphed in Fig. 6 below demonstrate different profiles of conversation, demonstrating that visitor groups can attain high scores through a variety of ratios of low/mid/high level talk.

In the study we present below, the average content score for all 119 coded sessions of groups using our interactive data visualization exhibit was 69.4 (*SD* = 42.6). Example A in Fig. 6 was a low-performing group, with a content score of 27. This interaction involved a largely one-sided dialogue, with one active participant narrating her activities and making some interpretive statements with very little input from her companion. Examples B, C, and D,









**Fig. 6** The SQuILD method is not intended to value a particular ratio as “best.” Instead, dialogue is evaluated based on how well it aligns to learning goals and how many “hooks” visitors have to engage with each other and the presented content

by contrast, all scored roughly half a standard deviation above the mean, but they achieved those scores in different ways. The pair in Example B had the highest proportion of high-relevance talk of any of our examples but overall fewer codes applied, resulting in a session score of 86. Examples C and D both had fewer high-relevance codings but made up for them with more low- and mid-relevance codes, resulting in scores of 92 and 94. By comparison, both members of the high-performing group in Example E were actively engaged in data interpretation, building off each other’s comments and their own observations. The richness of their discussion is evidenced by the high numbers of codes applied in all three categories and their high overall score of 185. The utility of this weighted coding system is that it allows different kinds of engagement (like Examples B, C, and D above) to be acknowledged as productive while still distinguishing low (e.g. Example A) and high (Example E) performing groups.

The final point to consider in applying a magnitude coding scheme is the numerical values assigned to each code level. The research team felt values of 1–2–3 for low-mid-high codes best reflected the relationship among code levels, but a full analysis vetted this assumption by testing two alternative scoring proportions (1–3–5 and 1–5–10). In the 119 visitor sessions analyzed for that study, the results of the A/B testing were consistent regardless of the scoring proportion, i.e. the same design “won” in all scoring, although some statistical significance was lost in the 1–5–10 proportion. The SQuILD methodology, by meaningfully segmenting dialogue and applying codes relevant to open-ended discussion and weighted according to their alignment with the exhibit’s learning goals, provides a valid quantitative measure for conducting A/B testing and informing exhibit design decisions.

## Employing SQuILD: An in situ study

Here we present an in situ  $2 \times 2$  experimental study exploring how changes to the control device for an interactive census data map exhibit (Fig. 1) impact visitors’ learning talk as measured by the SQuILD method. Specifically, this study varies the *means of control* (MoC) for the interactivity—whether visitors use a handheld tablet or a full-body interactive system to enact changes to the display—and the *distribution of control* (DoC) to a single user or multiple users. Visitors’ dialogue and interactions with the system are measured in order to compare conditions on their affordances for supporting productive learning talk (Fig. 7).

	Handheld (HH)	Full-body (FB)
Single-input (S)	 N = 33	 N = 29
Multi-input (M)	 N = 29	 N = 27

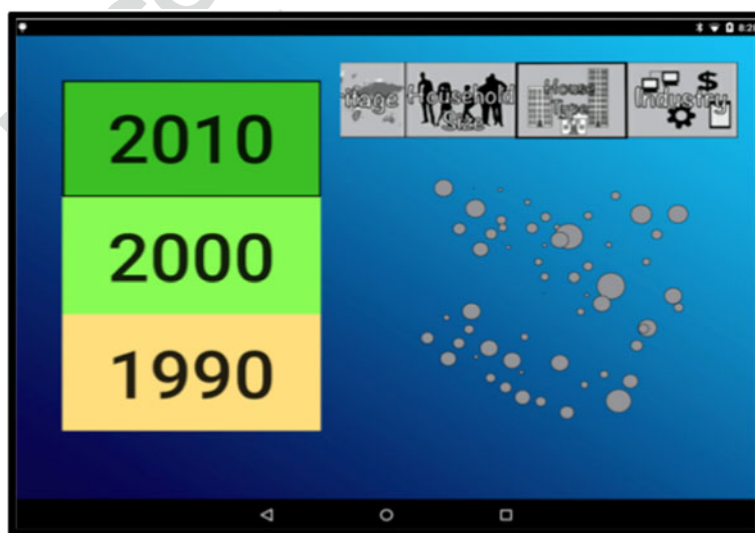
**Fig. 7** Two-by-two study design. 119 groups of two or more visitors participated in this study, with 57 groups using the full-body version and 62 in the handheld, and 57 groups using multi-input and 62 using single-input control

## Experimental conditions

The exhibit allows visitors to control three variables of the data visualization: decade, category, and aggregation level. Two form factors for enacting these control actions were tested. The handheld controller is a tablet-based interface that utilizes familiar touch controllers—buttons and sliders—to manipulate the visualization (see Fig. 8). The full-body controller condition places the locus of interaction and input in the user's body instead of an external device (Cafaro et al. 2014a, b).

## Interaction design

Throughout the course of this design-based research project, multiple control gestures have been tested in lab and museum settings, attending to embodied interaction concerns of semantic mappings between movements and gesture, internal and external consistency of gesture suites, and guessability (Cafaro, 2015; Cafaro et al. 2013; Roberts et al. 2014). The



**Fig. 8** Screen capture of tablet controller for HH conditions. Users manipulated the data display by tapping decades, swiping categories, or pinching aggregation levels

study presented here addresses an emergent question in this line of research – whether the size of a control action in an embodied interaction experience significantly impacts the users' outcomes. That is, if semantic mappings are held constant, will a physically bigger control action (i.e. using full-body control) support learning better than a smaller one? To test this, we developed a suite of whole-body gestures drawn from lab studies in this project (Cafaro, 2015; Cafaro et al. 2013) that have congruous tablet-based counterparts.

To select the census year of their data (1990, 2000, or 2010), visitors in the full-body condition step into marked areas on the floor, and visitors using the handheld controller push a button. Changing categories (e.g. ancestry, household size) is accomplished with a sideways swipe that in the handheld condition mimics the flick smart phone users utilize to browse pictures in a camera photo gallery, and in the full-body condition is the same gesture but bigger, as if the user is flipping through the pages of a very large book. Changing the aggregation level is accomplished by a pinch motion on the handheld and by pulling both hands apart or pushing them together in the full body condition, in a movement similar to that of playing the accordion.

In all cases two participants completed the kiosk survey in order to view their own self-selected data, but participants in the single-input conditions were given a single global controller that manipulated both data sets simultaneously, i.e. a single control action would change both datasets. Multi-input conditions allowed two participating visitors to independently control their own datasets via individual tablets (HH) or gestures (FB).

## Participants and research context

This study was conducted in situ in a museum using members of the general visitor population. All visitors interested in using the exhibit were invited to do so, and all visitor groups who orally consented to participate in the research and wear lapel microphones during their interactions were assigned a group ID number. The analysis presented here examines the 119 visitor sessions in which at least two visitors interacted together, spoke English during their interaction, and through their dialogue demonstrated that they understood what the data map represented.

The exhibit is situated in a small, irregularly shaped room approximately 2.9 m × 6 m separated from the museum's main hall by a partial wall. A 90" LCD flatscreen television serves as the main display at the far end of the room, away from the main entry door. In the full-body condition a rectangular "interaction area" is marked on the floor by a green floor mat labeled with the decades of data available for exploration (Fig. 1).

## Procedure

After pairs of participants completed their data profiles using the kiosk application (Fig. 2) they were given lanyards with a clip microphone. Visitors in the handheld conditions were given one or two Android tablets, depending on single- or multi-input condition, running the controller application. If more than two visitors were present in the group, spectators (those who did not complete the kiosk survey) were invited to watch from the back of the space where they could see the display and participate in the conversation. Museum staff explainers and members of the research team remained at the back of the room and only intervened in the interaction to answer questions. To ensure consistency in the responsiveness of the technology, we used a Wizard of Oz approach in the FB condition, where a member of the research team

was ready to manually trigger events in response to valid visitor gestures. Participants were allowed to remain in the interaction area as long as they wished. A video camera at the back of the room recorded visitor interactions, and the video feed was synchronized with microphone audio and a screen capture of the display for analysis.

### *Coding visitor sessions*

Sessions were segmented and coded for talk as described above. The raw counts of codes applied in each category and content scores from the magnitude coding were calculated for each session. We also computed the total interaction length, the number of data manipulations, the amount of time spent in silence, and the time spent learning to interact with the system (e.g. asking for instructions or help).

Seven sessions comprising 151 idea units were coded for dialogue by a second coder. In total, 467 codes were applied with an overall agreement of 85.65% (400 of 467 codes). Cohen's kappa was run and determined there was strong agreement,  $\kappa = .825, p < .001$ .

## **Findings**

### **Comparing conditions: Learning talk**

Visitors' learning talk was measured through analysis of the raw number of codes applied per session by condition, the content scores, and the rate of visitors' talk (i.e. score per second). The sections below identify statistical differences in these measures by the *means of control* (MoC: full-body or handheld) and *distribution of control* (DoC: single or multi-input).

### *Differences in number and categories of codes applied*

We first wanted to investigate if there were any characteristic differences in the nature of the talk across the conditions, before attempting to assign value to the talk using our weighting scheme. A two-way analysis of variance (ANOVA) yielded a main effect for MoC  $F(1115) = 5.841, p < .02$ , indicating that the average number of codes applied to a session was higher for handheld (HH) sessions ( $M = 56.61, SD = 29.23$ ) than full-body (FB) ( $M = 44.16, SD = 27.09$ ). The main effect for DoC was non-significant, as was the interaction effect. By category, a main effect was found for MoC on the number of *instantiate* and *integrate* codes applied, with the HH conditions receiving more of each type. In both of these categories, no main effect was found for DoC, nor was an interaction effect found. No significant main or interaction effects for MoC or DoC were found for *manage*, *evaluate*, or *generate* codes.

These findings suggest that the Means of Control somehow affected the character of visitors' conversations. The next section examines how those differences manifested in terms of the exhibit learning goals.

### *Differences in quality of learning talk*

To identify how well each condition was supporting the intended learning goals, each code was assigned a weight according to its relevance (Table 1). Then the weights were summed to

produce a content score for each session providing an overall picture of the quality of that session's learning talk. Content scores in this dataset ranged from 10 to 242, with a mean of 69.42 and a standard deviation of 42.60.

A two-way ANOVA on the scores yielded a main effect for MoC,  $F(1, 115) = 4.52$ ,  $p = .036$ , with a significantly higher weighted content score for HH sessions ( $M = 77.32$ ,  $SD = 43.93$ ) than FB ( $M = 60.82$ ,  $SD = 39.72$ ). The main effect for DoC was non-significant, as was the interaction effect. These findings reveal that, contrary to expectations, the handheld condition better supported learning talk among visitors.

### *Rate of learning talk*

Because the content scores represent a sum of all codes applied to a session's dialogue, differences in how long visitors spent interacting with the exhibit could affect their scores, i.e. visitors who linger longer have more time to talk and build up their scores. A one-way ANOVA was conducted to compare the four conditions on the duration of visitors' sessions. The ANOVA revealed that the effect of condition on session duration was significant,  $F(3,115) = 2.782$ ,  $p = .044$ . The handheld single-input (HHS) condition had the shortest average session duration at 145.8 s, compared to 191.9 s, 197.4 s, and 201.6 s for FBM, FBS, and HHM, respectively, but a Tukey HSD post-hoc comparison showed that no conditions were significantly different against any one other condition. However, when normalizing content scores by session duration to calculate the rate at which visitors produced substantive talk, significant effects were found.

The main effect of MoC yielded an F ratio of  $F(1,115) = 32.47$ ,  $p < .001$ , indicating that the normalized session score was significantly greater for HH conditions ( $M = .464$ ,  $SD = .160$ ) than for FB conditions ( $M = .311$ ,  $SD = .133$ ). The main effect of DoC yielded an F ratio of  $F(1,115) = 4.82$ ,  $p = .030$ , indicating that the normalized session score was significantly greater for single-input conditions ( $M = .421$ ,  $SD = .187$ ) than for multi-input conditions ( $M = .357$ ,  $SD = .132$ ). The interaction effect was significant,  $F(1,115) = 4.40$ ,  $p = .038$ , indicating that the MoC effect was greater in the single-input condition than the multi-input condition.

The combination of the higher content scores for HH conditions overall and the slightly shorter interaction times for the HHS condition led to the handheld single-input conditions producing the highest quality of data talk in the shortest amount of time, again contradicting expectations.

### **Comparing conditions: Interactions**

The measures of learning dialogue reported above indicate surprising findings that the handheld controller better supported visitor learning talk as measured by content scores, and did so especially efficiently when only one user had control. The next sections explore possible factors in the exhibit interactions that may have been able to account for these results.

### *How much data did visitors see?*

In this interactive exhibit, visitors were able to control not only what data were presented (via their selections at the kiosk), but also the amount of data rendered (i.e. shown on the display for at least 2 s) during their interactions due to the control actions they took. Upon stepping into the interaction area, two datasets were initially rendered: each visitor saw his selected



dataset for the first category (typically Heritage) in one decade (typically 2010). Any further rendering of data was only accomplished through control actions to switch the category and decades. Of the twenty-four potential data sets (4 categories  $\times$  3 decades  $\times$  2 visitors) multi-input (M) condition users could render any whole number between 2 and 24 datasets; for single-input conditions only even numbers of datasets (2, 4, ... 22, 24) could be rendered because the two sets were tied to each other.

Of the 119 sessions analyzed in this study, 44 (37%) rendered all 24 datasets. The minimum number rendered was 8 (5 groups). The mean was 20, with a standard deviation of 4.7. A moderate but significant correlation was found between the number of datasets rendered and the content score of a session,  $r(119) = .346, p < .001$ .

To relate this finding to the controller designs, a two-way ANOVA on the number of rendered datasets yielded a main effect for DoC,  $F(1, 115) = 4.14, p = .044$ , with a significantly higher number of rendered datasets for single-input sessions ( $M = 20.84, SD = 4.77$ ) than multi-input ( $M = 19.0, SD = 4.76$ ), which could possibly be explained by the “two-for-one” rendering attained by using the single-input controls. The main effect for MoC was non-significant,  $F(1, 115) = .043, p > .05$ , as was the interaction effect,  $F(1115) = 2.52, p > .05$ , so the higher-quality data talk in the HH condition is not explained by a higher degree of exposure to data.

Although differences in distribution of control were statistically significant, on average single-input groups saw less than two additional datasets ( $M = 20.84, SD = 4.77$ ) above their multi-input counterparts ( $M = 19.0, SD = 4.76$ ). This difference alone is unlikely to have greatly impacted the overall learning talk of the session, but it may have been partially responsible for differences among conditions. The next section looks more closely at engagement by examining differences in numbers of control actions.

### *Did handheld users make more moves?*

The two means of control tested here were both embodied in that they required physical movements in order to complete, but the smaller HH gestures may have been easier. We wondered whether movement was correlated with content scores, and if so, whether HH users made significantly more moves. Because movements were independent between users in the multi-input condition and linked in the single-input, we examine them separately here.

We first compared the overall number of control actions taken by each group. Independent samples t-tests were conducted to compare number of control moves for the FB and HH conditions. In the multi-input conditions, participants in the HHM condition made a higher number of control actions ( $M = 55.59, SD = 27.69$ ) than did those in the FBM condition ( $M = 39.00, SD = 14.38$ ),  $t(42.40) = -2.85, p = .007$ . A t-test on the single-input conditions indicated that no significant difference existed between HH ( $M = 32.24, SD = 20.88$ ) and FB ( $M = 33.59, SD = 19.73$ ) conditions,  $t(60) = .259, p > .05$ .

Session scores and total number of control moves were moderately positively correlated in the multi-input conditions,  $r(57) = .51, p < .01$  and were strongly correlated in single-input conditions,  $r(62) = .62, p < .001$ . In particular, a strong correlation existed between the number of decade moves and session score in the single-input condition,  $r(62) = .61, p < .001$ . Decade moves were positively but weakly correlated with session scores in the multi-input condition,  $r(57) = .27, p = .042$ .

These correlations suggest that physical interactions with an exhibit may provide an advantage for supporting learning talk. This finding would be in alignment with both theories

of the affordances of interactive data visualizations (e.g. Card et al. 1999) and of embodied cognition. However, this dataset does not support an analysis to tease apart which of these theories is more at play in this scenario; further work in this area is warranted.

If learning talk is correlated with control actions, the ease of completing those control actions is an important consideration for exhibit designers. While full-body interaction may have inherent affordances for cognition, those affordances may be negated by difficulties in completing the actions. Visitors were not directly surveyed about perceived difficulty of controlling the exhibit, so impacts of physical barriers cannot be measured using this dataset. The difficulty visitors had in understanding the control actions can be examined, however. A proxy for measuring this concern with the available data is presented next.

### *Did FB users have a harder time learning the controls?*

Any novel interactive system will require some amount of time for users to learn and become comfortable with the controls and basic operation of the interactive features. From the videos we marked out the time visitors spent talking to each other or the researchers about working the controls, which could affect the amount of productive data talk visitors could do and therefore impact visitors' session scores. In particular, we hypothesized that visitors in the FB conditions likely had a more difficult time learning the mechanics of the novel interaction than those in the HH conditions, and that multi-input users may have had to devote more talk to understand the control actions than single-input users, where one visitor in the interaction could "drive" the session. This measurement provides a proxy for assessing how difficult it was for visitors to understand how the exhibit worked.

To account for variation in average session duration, we computed "mechanical" time as a proportion of total session time. A two-way analysis of variance yielded a main effect for MoC,  $F(1115) = 52.55$ ,  $p < .001$ , indicating that users in the HH condition indeed devoted a smaller percentage of their interaction to figuring out the mechanics of the system ( $M = 7.94$ ,  $SD = 7.66$ ) than those in the full-body condition ( $M = 19.47$ ,  $SD = 9.50$ ). No significant DoC or interaction effects were found. If participants in the full-body interactive conditions spent on average 20% of their interaction time dealing with the mechanics of the system, it could follow that their content scores would be affected. Indeed, a significant but weak negative correlation was found between content scores and percentage of mechanical time across all sessions,  $r(119) = -.248$ ,  $p = .007$ .

Subtracting the mechanical time from the active session times can therefore give an adjusted metric for viewing session length. A two-way ANOVA comparing MoC and DoC on adjusted duration found the no significant differences on main effect of MoC, DoC, or their interaction, indicating that once the time spent talking about the mechanics of the system was subtracted from the amount of time spent interacting, all conditions spent a comparable amount of time during which they could have been engaging in productive learning talk. Rather than accounting for the variations in session scores, this finding makes these content score differences—particularly between FB and HH conditions—even more striking.

In fact, a two-way analysis of variance examining session scores normalized by the active, non-mechanical time (in seconds) yielded a main effect for MoC,  $F(1115) = 14.97$ ,  $p < .001$ , indicating that even when subtracting out the time spent learning the control of the system, visitors in the handheld sessions had a higher score-per-second ( $M = .50$ ,  $SD = .17$ ) than those in the full-body conditions ( $M = .39$ ,  $SD = .16$ ). The main effect for DoC was non-significant,

but the interaction effect was significant,  $F(1115) = 4.20$ ,  $p = .043$ , indicating that the MoC effect was greater in the single-input condition than in the multi-input condition.

These results suggest that it was not the time spent learning the control of the system that accounts for variations in learning talk among conditions. Implications for these findings are discussed next.

## Discussion

The field of CSCL has taken great interest in the role of the body in supporting collaborative learning (Davidsen and Ryberg 2017; Enyedy et al. 2012, 2015; Yoon et al. 2012). Advances in motion-sensing and augmented reality technologies have opened many doors for creating whole-body interactive computer supported learning environments in museums (Lyons 2016; Tscholl and Lindgren 2016; Uzzo et al. 2016; Yoon et al. 2012), but many methodological and theoretical questions remain. We have presented here a method, Scoring Qualitative Informal Learning Dialogue (SQuILD), that meets a need faced by many researchers examining intersubjective learning (Suthers 2006) in open-ended exploratory environments where individual cognitive assessments such as pre- and post-tests cannot fully capture the learning (Stahl, 2015). We then applied that methodology to an in situ study of an interactive museum exhibit, CoCensus, to find that our expectations were wholly contradicted by the results. In this discussion we will reflect on additional factors that may have been at play in the design of the exhibit, and we will conclude with considerations for employing this methodology in future studies.

## Framing embodiment

CoCensus is designed to support intersubjective learning (Suthers 2006) as users interact with data and each other. To encourage dialogue, the exhibit was conceptualized as a multi-user interactive exhibit drawing on theories of embodied cognition. Embodied interaction designs can influence learning via the embodiment, embedding, and extension of cognition. The CoCensus exhibit, like many other full-body interaction designs, contained elements of all three perspectives in its design. We held the *embodied* interaction design more or less constant across the tablet and the full body CoCensus designs: in both conditions, the visitor control of the data visualization was designed to be aligned with the same *embodied* schemata (moving forward and backward through time, scaling up and down as stretching or squashing, and selecting a new item by pushing aside the current item), although more sensorimotor action was required in the full body design than in the tablet design. With the full-body version of the CoCensus exhibit, we made sure that the physical movements were designed to be highly visible so as to *embed* the visitor actions more completely in the physical and social context to spur more collective dialogue. In the multi-user control versions of CoCensus, by allowing multiple learners to co-construct the state of the exhibit, learners should have been able to *extend* their reasoning about the exhibit to embrace both their and their companions' actions. By capitalizing on all three *embodied*, *embedded*, and *extended* aspects of situated cognition, the multi-user full-body condition should have more productively mediated learner interactions, but in the study presented here, this was not the case. Visitors in the full-body controller conditions produced significantly less learning talk as measured by content scores than those in the handheld conditions, whether in the single-input or multi-input variations. Analysis of

interactions attempted to account for these differences by examining differences in how people interacted with the exhibit in the conditions, including the amount of data rendered, the number of control actions made, and difficulties operating the exhibit. None of these factors were fully able to explain the results.

The CoCensus gesture suites for the full-body and handheld conditions were designed to be parallel to each other in an attempt to isolate the size of the gesture—the amount of sensorimotor engagement in the action—as the independent variable. However, this may have led to an unintended confound where full-body controls were less intuitive than their handheld counterparts. Participants in the handheld conditions of our study may have—by virtue of being handed a familiar tablet device—inadvertently been given a “frame” (Lakoff 2008) for how to interact with the system. They were able to employ a suite of familiar gestures commonly used with touchscreen devices: swiping, pinching, and button pressing. Perhaps more importantly, these gestures mapped clearly to familiar outcomes: swiping changed the “picture” of the map to another category just as swiping photos in a camera’s gallery application changes to a new picture, pinching the tablet controller was somewhat akin to a zoom that would be expected in a photo or map application (and in fact many visitors referred to aggregation changes as “zooming”), and tapping a decade button selected that decade as one would expect in any menu. Because these gestures are commonly used together as part of the suite of standard tablet gestures, successfully performing one of these gestures would allow a user to make inferences about which other gestures would be valid to use, and to seamlessly interleave their use of gestures. No comparable frame exists for the full-body equivalents of these gestures, although in our later work (Cafaro et al. 2014a, b) we pursued new methods for generating a unified suite of full-body actions that would seem to users to be part of normal life like those on a touchscreen, and preliminary results show that these suites do seem to be more discoverable.

## *Continuous gestures for discrete system responses*

A related issue with the embodied nature of CoCensus exhibit design is that the control actions in this exhibit activated discrete, not continuous, responses from the system, which may have interfered with visitors’ ability to fully embrace the full-body controls (Cafaro et al. 2014a, b). Completing any of the three control gestures triggered a change, but until the event was triggered the user received no feedback from small movements, potentially decreasing their ability to develop fluency of control (Norman, 2010; Snibbe and Raffle 2009). This is unlike full body systems in which the user’s sensorimotor manipulations are tightly coupled with system outputs, wherein the user can reflect on both proprioceptive information as well as system feedback to tune manipulations.

We have some evidence for this explanation being viable in the form of earlier design iterations of CoCensus (Roberts et al. 2014). In these early iterations, two types of visitor movement were directly and continuously linked to system responses: visitors’ distance from the display directly controlled the transparency of an individual’s data bubbles (fully opaque when the visitor was at the front of the room and nearly fully transparent at the back), and the speed of a visitor’s movements directly controlled the amount of animated jiggling of their data bubbles. The initial idea behind altering the transparency was that it would allow visitors to focus on either the data or the underlying map geography by making one or the other more visible. The bubble jiggling effect was present to help the visitors understand who was controlling each data set. These interaction components were abandoned following findings

that the front-to-back timeline configuration was more intuitive and enjoyable for users (Roberts et al. 2014), so movement perpendicular to the display became the timeline controller rather than a transparency controller. Because the timeline made clear which visitor was controlling which data set, we also removed the jiggling effect. We now suspect that these earlier design decisions had the side effect of fostering stronger physical and conceptual links between the person and data. Interaction design utilizing continuous feedback likely takes better advantage of the affordances of full-body interaction, suggesting that comparing discrete versus continuous full-body controls would be valuable future work.

## Conclusions

This application of the SQuILD method reveals the value of using quantitative measures of learning talk in an open-ended, collaborative museum exhibit to force researchers to confront their assumptions about theory-grounded designs. Via our methodology, we discovered that full-body interaction designs do not always perform as theory might predict, and brought to light several intersecting design considerations and areas for future research. In interpreting these findings, it is important first to recognize two caveats. (1) This research was structured as an experiment where the measured outcome was learning talk, so we did not dive into the mechanics of how their physical actions may have affected their social embedding. It could be valuable to examine their actions and talk to determine the extent to which visitors used one another's actions as semiotic resources. (2) This research was structured using a sociocultural framework, meaning that we did not collect individual-level measures that could speak to how the physical embodiment affected individual users. For example, we did not collect affective measures, such as visitors' enjoyment of the system. This metric could be studied qualitatively through exclamations and other affective responses (Allen 2002), and may show a preference of museum visitors for more highly physical interactions.

One mother in the HHS condition, for example, said sarcastically as her son was handed the tablet controller, "Wow, I'm so glad we came to the museum today to get you away from your iPad," and some participants in the full-body conditions—particularly younger users—clearly enjoyed jumping around the timeline. Measures of individual cognitive gain could also demonstrate affordances of full-body interaction on cognition that may not have manifested in dialogue. Similarly, this analysis did not look at individual interactions to measure "participation equity" (Kapur and Kinzer 2007) or "task division" (Lyons 2009). An analysis narrowing the focus to individuals' contributions within sessions, particularly in relation to their movements and gestures, could deepen our knowledge of how movement mediated individuals' experiences.

These potential future lines of research point to the limitations of the SQuILD methodology. Specifically, SQuILD does not replace other methodologies for understanding how people learn – in fact, it does not help us understand the *process* of intersubjective meaning-making, which has been proposed as a goal for CSCL methods (Suthers 2006). By distilling rich talk into a single number, SQuILD glosses over key nuances in dialogue and obfuscates the process of knowledge building among group members over the course of an interaction. Instead, SQuILD fills another important need. As the possibilities for CSCL expand off the desktop and into user-driven, open exploration contexts, the methodologies often used to study learning (e.g., individual measures like pre- and post-tests or coding of individual utterances) need to be



re-examined and refined (Stahl, 2015). Here we have argued that empirical measurements and statistical analyses can reliably be applied to the evaluation of open-ended group collaborative learning. This methodology permits direct comparisons of competing designs in intersubjective CSCL studies focusing on supporting learning talk, providing empirical results to guide ongoing design.

It is worth noting in the example study presented here that even after having completed the dialogue coding, the research team still expected the full-body multi-input condition to have been the most successful in supporting learning talk. It was our favorite, it was grounded in exciting theories and promising lab studies, and we'd seen enough compellingly successful FBM sessions in situ to allow confirmation bias to set in. It was only after the statistical tests were run (and re-run and double-checked, and the coding triple-checked) that we were able to admit that our assumptions about the design's affordances didn't play out in the authentic museum context. Researchers can usually find exemplars of how a new design meets expectations if they look hard enough – but these cherry-picked examples don't speak to how *consistently* the design works as intended. We hope that this methodology—and additional work building on and refining it for broader applications—will similarly allow researchers to question their assumptions as they assess the ecological validity of their design decisions in CSCL research.

**Acknowledgements** This material is based upon work supported by the National Science Foundation under Grant No. 1248052.

## References

- Abrahamson, D., & Bakker, A. (2016). Making sense of movement in embodied design for mathematics learning. *Cognitive Research: Principles and Implications*, 1, 33.
- Abrahamson, D., & Sánchez-García, R. (2016). Learning is moving in new ways: The ecological dynamics of mathematics education. *Journal of the Learning Sciences*, 25(2), 203–239.
- Allen, S. (2002). Looking for learning in visitor talk: A methodological exploration. In *Learning conversations in museums* (pp. 259–303).
- Allen, S. (2004). Designs for learning: Studying science museum exhibits that do more than entertain. *Science Education*, 88(S1), S17–S33. <https://doi.org/10.1002/sce.20016>.
- Angelo, S. D., Pollock, D. H., & Horn, M. (2015). Fishing with friends : Using tabletop games to raise environmental awareness in aquariums. In *Proceedings of IDC 2015* (pp. 29–38).
- Antle, A. N., Droumeva, M., & Corness, G. (2008). Playing with the sound maker: Do embodied metaphors help children learn? In *Proceedings of IDC '08* (pp. 178–185). ACM Press.
- Antle, A. N., Corness, G., & Droumeva, M. (2009). What the body knows: Exploring the benefits of embodied metaphors in hybrid physical digital environments. *Interacting with Computers*, 21(1), 66–75.
- Antle, A. N., Corness, G., & Bevans, A. (2013a). Balancing justice: Comparing whole body and controller-based interaction for an abstract domain. *International Journal of Arts and Technology*, 6(4), 388.
- Antle, A. N., Wise, A. F., Hall, A., Nowroozi, S., Tan, P., Warren, J., ... Fan, M. (2013b). Youtopia. In *Proceedings of IDC 2013* (pp. 565–568). ACM Press.
- Aoki, P. M., Grinter, R. E., Hurst, A., Szymanski, M. H., Thornton, J. D., & Woodruff, A. (2002). Sotto voce: Exploring the interplay of conversation and mobile audio spaces. In *Proceedings of CHI 2002* (pp. 431–438). ACM Press.
- Ash, D. (2003). Dialogic inquiry in life science conversations of family groups in a museum. *Journal of Research in Science Teaching*, 40(2), 138–162.
- Ash, D. (2004). How families use questions at dioramas: Ideas for exhibit design. *Curator: The Museum Journal*, 47(1), 84–100.
- Atkins, L. J., Velez, L., Goudy, D., & Dunbar, K. N. (2009). The unintended effects of interactive objects and labels in the science museum. *Science Education*, 93(1), 161–184.
- Beheshti, E., Obiorah, M., & Horn, M. S. (2015). “Let’s dive into it!”: Learning electricity with multiple representations. In *Proceedings of IDC 2015* (pp. 263–266). ACM Press.

- Bernard, H. R. (2006). *Research methods in anthropology: Qualitative & quantitative approaches*. Lanham: AltaMira Press. 1122
- Bitgood, S. (2000). The role of attention in designing effective interpretive labels. *Journal of Interpretation Research*, 5(2), 31–45. 1123
- Block, F., Hammerman, J., Horn, M., Spiegel, A., Christiansen, J., Phillips, B., ... & Shen, C. (2015). Fluid grouping: Quantifying group engagement around interactive tabletop exhibits in the wild. In *Proceedings of CHI 2015* (pp. 867–876). ACM Press. 1124
- Borun, M., Chambers, M., & Cleghorn, A. (1996). Families are learning in science museums. *Curator*, 39(2), 262–270. 1125
- Cabrera, J. S., Frutos, H. M., Stoica, A. G., Avouris, N., Dimitriadis, Y., Fiotakis, G., & Liveri, K. D. (2005). Mystery in the museum: Collaborative learning activities using handheld devices. In *Proceedings of MobileHCI 2005* (pp. 315–318). Salzburg, Austria: ACM Press. 1126
- Cafaro, F., Panella, A., Lyons, L., Roberts, J., & Radinsky, J. (2013) I see you there! Developing identity-preserving embodied interaction for museum exhibits. In *Proceedings of CHI 2013* (pp. 1911–1920). ACM Press. 1127
- Cafaro, F., Lyons, L., Kang, R., Radinsky, J., Roberts, J., Vogt, K. F. (2014a). Framed Guessability: Using embodied allegories to increase user agreement on gesture sets. In *Proceedings of TEI 2014* (pp. 197–204). ACM Press. 1128
- Cafaro, F., Lyons, L., Roberts, J., & Radinsky, J. (2014b). The uncanny valley of embodied interaction design. In *Proceedings of DIS 2014* (pp. 1075–1078). ACM Press. 1129
- Card, S., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufman. 1130
- Charoenying, T. (2013). Graph hopping: Learning through physical interaction quantification. In *Proceedings of IDC 2013* (pp. 495–498). ACM Press. 1131
- Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, 6(3), 271–315. 1132
- Crowley, K., & Jacobs, M. (2002). Building islands of expertise in everyday family activity. In G. Leinhardt, K. J. Crowley, & K. Knutson (Eds.), *Learning conversations in museums* (pp. 333–356). Mahwah: Lawrence Erlbaum. 1133
- Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, 18, 382–393. 1134
- Davidsen, J., & Ryberg, T. (2017). “This is the size of one meter”: Children’s bodily-material collaboration. *International Journal of Computer-Supported Collaborative Learning*, 12(1), 65–90. 1135
- Davis, P., Horn, M., Block, F., Phillips, B., Evans, E. M., Diamond, J., & Shen, C. (2015). “Whoa! We’re going deep in the trees!”: Patterns of collaboration around an interactive information visualization exhibit. *International Journal of Computer-Supported Collaborative Learning*, 10(1), 53–76. 1136
- Diamond, J., Bond, A., Schenker, B., Meier, D., & Twersky, D. (1995). Collaborative multimedia. *Curator*, 38(3), 137–149. 1137
- Diamond, J., Luke, J. J., & Uttal, D. H. (2009). *Practical evaluation guide: Tools for museums & other informal educational settings* (2 ed.). Lanham: AltaMira Press. 1138
- Dini, R., Patemò, F., & Santoro, C. (2007). An environment to support multi-user interaction and cooperation for improving museum visits through games. In *Proceedings of MobileHCI '07* (pp. 515–521). ACM press. 1139
- Dourish, P. (2001). *Where the action is : The foundations of embodied interaction*. Cambridge: MIT Press. 1140
- Eberbach, C., & Crowley, K. (2005). From living to virtual: Learning from museum objects. *Curator*, 48(3), 317–338. 1141
- Engeström, Y., Miettinen, R., & Punamäki-Gitai, R.-L. (1999). *Perspectives on activity theory*. Cambridge: Cambridge University Press. 1142
- Enyedy, N., Danish, J. A., Delacruz, G., & Kumar, M. (2012). Learning physics through play in an augmented reality environment. *International Journal of Computer-Supported Collaborative Learning*, 7(3), 347–378. 1143
- Enyedy, N., Danish, J. A., & DeLiema, D. (2015). Constructing liminal blends in a collaborative augmented-reality learning environment. *International Journal of Computer-Supported Collaborative Learning*, 10(1), 7–34. 1144
- Falcão, T. P., & Price, S. (2009). What have you done! The role of ‘interference’ in tangible environments for supporting collaborative learning. In *Proceedings of CSCL 2009* (pp. 324–334). ISLS. 1145
- Falk, J. H. (2006). An identity-centered approach to understanding museum learning. *Curator*, 49(2), 151–166. 1146
- Falk, J. H., & Dierking, L. D. (2000). *Learning from museums : Visitor experiences and the making of meaning*. American Association for State and Local History book series. Walnut Creek: AltaMira Press. 1147
- Falk, J. H., & Storksdieck, M. (2005). Using the contextual model of learning to understand visitor learning from a science center exhibition. *Science Education*, 89(5), 744–778. 1148

- Fischer, U., Moeller, K., Bientzle, M., Cress, U., & Nuerk, H.-C. (2011). Sensori-motor spatial training of number magnitude representation. *Psychonomic Bulletin & Review*, 18(1), 177–183. 1181
- Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2), 124–158. 1182
- Gallagher, S., & Lindgren, R. (2015). Enactive metaphors: Learning through full-body engagement. *Educational Psychology Review*, 27(3), 391–404. 1183
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin. 1184
- Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10), 1489–1522. 1185
- Heath, C., & vom Lehn, D. (2008). Configuring 'Interactivity' enhancing engagement in science centres and museums. *Social Studies of Science*, 38(1), 63–91. 1186
- Hindmarsh, J., Heath, C., Vom Lehn, D., & Cleverly, J. (2005). Creating assemblies in public environments: Social interaction, interactive exhibits and CSCW. *Computer Supported Cooperative Work (CSCW)*, 14(1), 1–41. 1187
- Hope, T., Nakamura, Y., Takahashi, T., Nobayashi, A., Fukuoka, S., Hamasaki, M., & Nishimura, T. (2009). Familial collaborations in a museum. In *Proceedings of CHI 2009* (pp. 1963–1972). ACM Press. 1188
- Horn, M., Roberts, J., Banerjee, A., & McGee, S. (2017) Touch | Don't touch: Exploring the role of interactive displays in natural history museums to help visitors appreciate objects behind glass. In *Proceedings of CSCW 2017* (pp. 851–852). ACM Press. 1189
- Howison, M., Trninic, D., Reinholz, D., & Abrahamson, D. (2011). The mathematical imagery trainer: From embodied interaction to conceptual learning. In *Proceedings of CHI 2011* (pp. 1989–1998). ACM Press. 1190
- Humphrey, T., Gutwill, J., & Exploratorium APE Team. (2005). *Fostering active prolonged engagement: The art of creating APE exhibits*. San Francisco: Exploratorium. 1191
- Jacobs, J. K., Yoshida, M., Stigler, J. W., & Fernandez, C. (1997). Japanese and American teachers' evaluations of mathematics lessons: A new technique for exploring beliefs. *The Journal of Mathematical Behavior*, 16(1), 7–24. 1192
- Johnson-Glenberg, M. C., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality: How gesture and motion capture affect physics education. *Cognitive Research: Principles and Implications*, 2(1), 24. 1193
- Johnson-Glenberg, M. C., Birchfield, D. A., Tolentino, L., & Koziupa, T. (2014). Collaborative embodied learning in mixed reality motion-capture environments: Two science studies. *Journal of Educational Psychology*, 106(1), 86–104. 1194
- Kaptelinin, V. (2013). The mediational perspective on digital technology: Understanding the interplay between technology, mind and action. In S. Price, C. Jewitt, & B. Brown (Eds.), *The SAGE Handbook of Digital Technology Research* (pp. 203–213). 1195
- Kapur, M., & Kinzer, C. K. (2007). Examining the effect of problem type in a synchronous computer-supported collaborative learning (CSCW) environment. *Educational Technology Research and Development*, 55(5), 439–459. 1196
- Kaschak, M. P., Connor, C. M., Dombek, J. L., Glenberg, A., Aveyard, M., & Blanchard, A. (2017). Enacted reading comprehension: Using bodily movement to aid the comprehension of abstract text content. *PLoS One*, 12(1), e0169711. 1197
- Kay, M., Kola, T., Hullman, J., & Munson, S. (2016). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings CHI 2016* (pp. 5092–5103). ACM Press. 1198
- Kisiel, J., Rowe, S., Vartabedian, M. A., & Kopczak, C. (2012). Evidence for family engagement in scientific reasoning at interactive animal exhibits. *Science Education*, 96(6), 1047–1070. 1199
- Kruppa, M., & Aslan, I. (2005). Parallel presentations for Heterogenous user groups - an initial user study. In M. Maybury et al. (Eds.), *Proceedings INTETAIN 2005, Lecture Notes in Computer Science* (pp. 54–63). Berlin: Springer. 1200
- Lakoff, G. (2008). The neural theory of metaphor. In R. W. Gibbs (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 17–38). Cambridge: Cambridge University Press. 1201
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press. 1202
- Lanir, J., Wecker, A. J., Kuflik, T., & Felberbaum, Y. (2016). Shared mobile displays: An exploratory study of their use in a museum setting. *Personal and Ubiquitous Computing*, 20(4), 635–651. 1203
- Lee, V. R., & Drake, J. (2013). Quantified recess: Design of an activity for elementary students involving analyses of their own movement data. In *Proceedings of IDC 2013* (pp. 273–276). ACM Press. 1204
- Leinhardt, G., & Crowley, K. (1998). Conversational elaboration as a process and an outcome of museum learning. In *Museum learning collaborative technical report (MLC-01)*. Pittsburgh: Learning Research and Development Center, University of Pittsburgh. 1205

- Leinhardt, G., & Knutson, K. (2004). *Listening in on museum conversations*. Walnut Creek: AltaMira Press. 1240
- Leinhardt, G., Crowley, K., & Knutson, K. (2002). *Learning conversations in museums*. Mahwah: Lawrence Erlbaum. 1241
- Liccardi, I., Abdul-Rahman, A., & Chen, M. (2016). I know where you live: Inferring details of people's lives by visualizing publicly shared location data. In *Proceedings of CHI 2016* (pp. 1–12). ACM Press. 1242
- Lindgren, R., & Johnson-Glenberg, M. (2013). Emboldened by embodiment: Six precepts for research on embodied learning and mixed reality. *Educational Researcher*, 42(8), 445–452. 1243
- Lindgren, R., Tscholl, M., Wang, S., & Johnson, E. (2016). Enhancing learning and engagement through embodied interaction within a mixed reality simulation. *Computers & Education*, 95, 174–187. 1244
- Lyons, L. (2009). Designing opportunistic user interfaces to support a collaborative museum exhibit. In *Proceedings of CSCL 2009* (pp. 375–384). ISLS. 1245
- Lyons, L. (2016). Exhibiting data: Using body-as-interface designs to engage visitors with data visualizations. In V. Lee (Ed.), *Learning technologies and the body: Integration and implementation in formal and informal learning environments*. New York: Routledge (Taylor & Francis). 1246
- Lyons, L., Slattery, B., Jimenez Pazmino, P., Lopez Silva, B., & Moher, T. (2012). Don't forget about the sweat: Effortful embodied interaction in support of learning. In *Proceedings TEI 2012* (pp. 77–84). ACM Press. 1247
- Lyons, L., Tissenbaum, M., Berland, M., Eydt, R., Wielgus, L., & Mechtley, A. (2015). Designing visible engineering: Supporting tinkering performances in museums. In *Proceedings of IDC 2015* (pp. 49–58). ACM Press. 1248
- Macedonia, M. (2003). Revitalizing museums with digital technology. *Computer*, 36(2), 94–96. 1249
- Malinverni, L., & Burgués, N. P. (2015). The medium matters: The impact of full-body interaction on the socio-affective aspects of collaboration. In *Proceedings of IDC 2015* (pp. 89–98). ACM Press. 1250
- Malinverni, L., Ackermann, E., & Pares, N. (2016). Experience as an object to think with. In *Proceedings of TEI 2016* (pp. 332–339). ACM press. 1251
- McLean, K. (1999). Museum exhibitions and the dynamics of dialogue. *Daedalus*, 128(3), 83–107. 1252
- Meisner, R., vom Lehn, D., Heath, C., Burch, A., Gammon, B., & Reisman, M. (2007). Exhibiting performance: Co-participation in science centres and museums. *International Journal of Science Education*, 29(12), 1531–1555. 1253
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks: Sage. 1254
- Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1998). *Adult literacy in OECD countries: Technical report on the first international adult literacy survey*. SSOP, Washington, DC: US Government Printing Office, Superintendent of Documents, Mail Stop. 1255
- National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, DC: National Academies Press. 1256
- Norman, D. (1988). *The psychology of everyday things*. New York: Basic Books. 1257
- Peppler, K., Danish, J., Zaitlen, B., Glosson, D., Jacobs, A., & Phelps, D. (2010). BeeSim: Leveraging wearable computers in participatory simulations with young children. In *Proceedings of IDC 2010* (pp. 246–249). ACM Press. 1258
- Povis, K. T., & Crowley, K. (2015). Family learning in object-based museums: The role of joint attention. *Visitor Studies*, 18(2), 168–182. 1259
- Robbins, P., & Aydede, M. (2009). A short primer on situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 3–10). Cambridge: Cambridge University Press. 1260
- Roberts, L. C. (1997). *From knowledge to narrative: Educators and the changing museum*. Washington, DC: Smithsonian Institution Press. 1261
- Roberts, J., Lyons, L., Cafaro, F., & Eydt, R. (2014). Interpreting data from within: Supporting human-data interaction in museum exhibits through perspective taking. In *Proceedings of IDC 2014* (pp. 7–16). ACM Press. 1262
- Roberts, J., Lyons, L., Cafaro, F., & Eydt, R. (2015). Harnessing motion-sensing technologies to engage visitors with digital data. In *Proceedings of Museums and the Web*. Archives & Museum Informatics. 1263
- Rounds, J. J. (2006). Doing identity work in museums. *Curator*, 49(2), 133–150. 1264
- Saldaña, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks: Sage. 1265
- Schauble, L., Gleason, M., Lehrer, R., Bartlett, K., Petrosino, A., Allen, A., et al. (2002). Supporting science learning in museums. In G. Leinhardt, K. Crowley, & K. Knutson (Eds.), *Learning conversations in museums* (pp. 333–356). Mahwah: Lawrence Erlbaum. 1266
- Scheible, J., & Ojala, T. (2005). MobiLenin combining a multi-track music video, personal mobile phones and a public display into multi-user interactive entertainment. In *Proceedings of MULTIMEDIA 2005* (pp. 199–208). ACM Press. 1267

- Schmitt, B., Bach, C., Dubois, E., & Duranthon, F. (2010). Designing and evaluating advanced interactive experiences to increase visitor's stimulation in a museum. In *Proceedings of the Augmented Human International Conference* (pp. 1–8). ACM Press.
- Seifert, C., & Hutchins, E. (1992). Error as opportunity: Learning in a cooperative task. *Human-Computer Interaction*, 7(4), 409–435.
- Serrell, B. (1996). *Exhibit labels: An interpretive approach*. Walnut Creek: Altamira Press.
- Simon, N. (2010). *The participatory museum*. Santa Cruz: Museum 2.0.
- Slattery, B., Lyons, L., Jimenez Pazmino, P., Lopez Silva, B., & Moher, T. (2014). How interpreters make use of technological supports in an interactive zoo exhibit. In *Proceedings of ICLS 2014* (pp. 198–205). ACM Press.
- Snibbe, S., & Raffle, H. (2009). Social immersive media. In *Proceedings of CHI 2009* (pp. 1447–1456). ACM Press.
- Steier, R. (2014). Posing the question: Visitor posing as embodied interpretation in an art museum. *Mind, Culture, and Activity*, 21(2), 148–170.
- Steier, R., Pierroux, P., & Krange, I. (2015). Embodied interpretation: Gesture, social interaction, and meaning making in a national art museum. *Learning, Culture and Social Interaction*, 7, 28–42.
- Suchman, L. (1987). *Plans and situated actions : The problem of human-machine communication*. New York: Cambridge University Press.
- Suthers, D. D. (2006). Technology affordances for intersubjective meaning making: A research agenda for CSCS. *International Journal of Computer-Supported Collaborative Learning*, 1(3), 315–337.
- Tashakkori, A., & Teddlie, C. (2010). *Sage handbook of mixed methods in social & behavioral research* (2nd ed.). Los Angeles: SAGE Publications.
- Tscholl, M., & Lindgren, R. (2016). Designing for learning conversations: How parents support Children's science learning within an immersive simulation. *Science Education*, 100(5), 877–902.
- Uzzo, S. M., Chen, R. S., & Downs, R. R. (2016). Connected Worlds: Connecting the public with complex environmental systems. In *American Geophysical Union, Fall General Assembly 2016, abstract #ED23F-05*.
- vom Lehn, D., Heath, C., & Hindmarsh, J. (2001). Conduct and collaboration in museums and galleries. *Symbolic Interaction*, 24(2), 189–216.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. Cambridge: Harvard University Press.
- Wagner, D., Schmalstieg, D., & Billinghurst, M. (2006). Handheld AR for collaborative edutainment. In *Proceedings of ICAT 2006* (pp. 85–96). Springer-Verlag.
- Wertsch, J. V. (1994). The primacy of mediated action in sociocultural studies. *Mind, Culture, and Activity*, 1(4), 202–208.
- Wertsch, J. V. (1998). *Mind as action*. USA: Oxford University Press.
- Williams, A., Kabisch, E., & Dourish, P. (2005). From interaction to participation: Configuring space through embodied interaction. In *Proceedings of Ubicomp 2005* (pp. 287–304). ACM Press.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636.
- Yap, K., Zheng, C., Tay, A., Yen, C.-C., & Do, E. Y.-L. (2015). Word out! In *Proceedings of the 6th Augmented Human International Conference on - AH '15* (pp. 101–108). ACM Press.
- Yatani, K., Sugimoto, M., & Kusunoki, F. (2004). Musex: A system for supporting children's collaborative learning in a museum with PDAs. In *Proceedings of WMUTE 2004* (pp. 109–112). IEEE.
- Yoon, S. A., & Wang, J. (2014). Making the invisible visible in science museums through augmented reality devices. *TechTrends*, 58(1), 49–55.
- Yoon, S. A., Elinich, K., Wang, J., Steinmeier, C., & Tucker, S. (2012). Using augmented reality and knowledge-building scaffolds to improve learning in a science museum. *International Journal of Computer-Supported Collaborative Learning*, 7(4), 519–541.
- Zhu, Y. (2007). Measuring effective data visualization. In G. Bebis, R. Boyle, B. Parvin, et al. (Eds.), *Advances in visual computing* (pp. 652–661). Berlin: Springer.