Scoring Qualitative Informal Learning Dialogue: The SQuILD Method for Measuring Museum Learning Talk

Jessica Roberts, University of Illinois at Chicago, jessicaannroberts@gmail.com
Leilah Lyons, University of Illinois at Chicago and New York Hall of Science, llyons@uic.edu

Abstract: Museums are increasingly developing computer-supported collaborative learning experiences and are in need of methods for evaluating the educational value of such exhibits. Exhibit designers, like web interaction designers, have long been employing A/B testing of exhibit elements in order to understand the affordances of competing designs *in situ*. When the exhibit elements being tested are intended to support open-ended group exploration and dialogue, existing knowledge-based metrics like measuring the amount of content recalled don't quite apply. Visitor groups can explore the educational content in idiosyncratic ways, meaning that not all groups have the same exposure to content, and the learning arises from the visitors' conversations. In order to evaluate learning outcomes for CSCL exhibits, we present a method for quantifying idiosyncratic social learning, Scoring Qualitative Informal Learning Dialogue (SQuILD), and demonstrate how it was applied to A/B testing of a collaborative data visualization exhibit in a metropolitan museum.

Introduction

Interaction designers have long relied on A/B testing to compare competing designs in real settings of use on measurable outcomes like duration of interaction and number of events logged during a session. These measures give valuable information about the use and benefits of interactive systems, but they are typically most informative for single user applications like websites. In multi-user systems such as collaborative interactive museum exhibits, interpersonal interactions are typically key aspects of the experience, yet have been traditionally difficult to quantify (Block et al., 2015). Museum designers have used A/B tests to evaluate exhibit elements like alternate versions of label designs (Serrel, 1996), using metrics like dwell time or information recall (Bitgood, 2000), but an increasing body of literature studying learning in museums suggests that such assessments miss the full breadth of learning occurring in these settings.

Instead, recent work has attended more closely to the role dialogue plays in informal learning. 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). Visitors' dialogue during their interactions with an exhibit is thus presumed to be the primary vehicle for learning, 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. This paper describes a methodology that solves three key problems in quantifying data 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.

Background

Challenge 1: Identifying open-ended learning with interactive exhibits

Museums are increasingly using novel interactive exhibits to create engaging spaces where visitors can shape their own narratives of the experience (Roberts, 1997). Unfortunately for exhibit designers and learning researchers, it is impossible to design a knowledge-based post-test to demonstrate content mastery for an open-ended exhibit, in which we have no expectations that all visitors will explore the same content. Other measures like visitor 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 & vom Lehn, 2008). Studying the dialogue of visitors while they are engaged with the exhibit may be more reliable, and researchers have made gains toward understanding what is productive talk given the free-choice environment, 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 & Knutson, 2004; Kisiel, Rowe, Vartabedian, & Kopcek, 2012). The foundation laid by this prior work, largely developed before exhibits were highly-interactive, needs to be built upon to address challenges specific to highly interactive computer-based exhibits, where visitors have increasing agency in shaping their own interaction experiences.

Challenge 2: Segmenting dialogue to permit cross-group comparisons

The grain size for segmentation is a key decision in any analytical process. Due to the spontaneous nature of joint exploration of a museum exhibit, many ideas are split among two or more visitors as they work together to make sense of the data. 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 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. Some meaningful segmentation is necessary to measure and compare session dialogue. A common segmentation strategy 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 would 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.

Challenge 3: Respecting socially-constructed learning when quantifying dialogue

The ultimate goal of the analysis described here is to create a valid quantitative measure of spontaneous dialogue during exhibit interactions for the purpose of conducting A/B testing of multiple exhibit designs, e.g. competing form factors for control of an interactive exhibit (Roberts & Lyons, in progress). As noted above, challenges in segmenting dialogue already make quantification of talk difficult, and complicating the process even further 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 fall short of designers' goals for the exhibit. For example, it is widely accepted that reading a label aloud is a productive form of talk in museums (Borun, Chambers, & Cleghorn, 1996; Kisiel et al., 2012; Allen, 2002; Atkins et al., 2009), but should such reading merit the same quantitative score as, for example, a comparison of two datasets, given that the aim of the exhibit is to foster such comparisons? But conversely, should a "conversation" where the only talk is one visitor making a single comparison be valued more highly than the extended conversation of a group that involves the reading aloud of a lot of low-level exhibit content and linking of that content to personal experiences? A productive 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.

SQuILD: A method for quantitatively comparing informal learning talk

Here we present the SQuILD (Scoring Qualitative Informal Learning Dialogue) method, and we provide examples of applying it to an interactive data map museum exhibit.

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

We began by identifying common threads in the literature on visitor talk to develop five categories of substantive talk—management, instantiations, evaluations, integrations, and generations. We then used an open coding process, informed by the literature of our exhibit's content domain (graph interpretation), to determine specific sub-codes within each of those categories (see Table 1). While our sub-codes may not be directly applicable to an exhibit with different content, the five categories could easily 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

It is to be expected that when multiple people are interacting with an exhibit, particularly when that exhibit is based on a novel technology, some amount of talk will directly address the interaction with the exhibit. Talk that 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 to researchers of museum learning 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 & Crowley, 2005, Diamond et al., 1986; Atkins et al., 2009). Researchers of technology-based multi-user interactives are similarly concerned with interpersonal interactions like interference (Falcão & Price, 2009), negotiation of exploration (Davis et al., 2013), and collaboration (Williams, Kabisch, & Dourish, 2005).

Instantiate codes

Here the term "instantiation" indicates when a user makes information part of the conversation by saying it aloud. The instantiation of information provides opportunities for the individual visitors to internalize that information (i.e., learn from the exhibit) and can lay the foundation for further reasoning among learners on a museum visit (Kisiel, et al., 2012). Saying something aloud is an important part of the social learning process: putting ideas into the shared social space and helping establish joint attention (also referred to as "grounding"). Per sociocultural learning theory, learners must articulate ideas via communication before learning can take place (Vygotsky, 1978). Processes of noticing and establishing joint attention among visitor group members have been found to be productive in facilitating learning talk in museums (Povis & Crowley, 2015; Leinhardt & 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 go beyond merely instantiating content to make some kind of judgment or assessment about a piece of information by assigning some kind of value, whether qualitative or quantitative. Such personal qualitative evaluations are arguably 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 or part of a more complex statement. The most common sub-code of evaluative statement in our exhibit was *characterize*. Examples of *characterize* evaluation statements are those remarking that there are "a lot" or "not very many" of something, or describing a population as being "everywhere." The characterizations could be spatial or quantitative in nature.

Integrate codes

While evaluation statements refer to a single idea, the final two categories connect multiple pieces of information in some way. Friel, Curcio, and Bright (2001) refer to the act of looking for relationships in data as "interpretation." The SQuILD coding framework adopts the more precise term *integration* from Murray, Kirsch, & Jenkins (1997) 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 & 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 & Dierking's (2000) Contextual Model of Learning posits that what learners gain during a learning experience is inextricably tied to what 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."

Addressing challenge 2: Segmenting dialogue through idea units

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* 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 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?	[draw joint attention to areas that changed over time]
[3]	B:	Yeah.	
[4]	A:	And up there.	
[5]	B:	More spread out.	[characterize data]
[6]	A:	But you see the greatest change here on this side.	[identify area of particular interest]

This excerpt was divided into four idea units. These idea units 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. 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 a dashed 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 Figure 1). 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.

Figure 1. Idea units that overlap each other in time can be segmented in video, shown here as white bars in the coding software MaxQDA.

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. In our work, any idea units that did not match any of the above categories and sub-codes were marked *non-substantive* and were disregarded in the analysis. Some idea units were coded with a 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, consider the statement:

"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 INSTANTIATE-compare—would give it the same value as a much less rich statement like, "It looks like there are more of them." Only by 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 stays close to the data in identifying how people are talking by flagging 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 & Huberman, 1994) by sorting the sub-codes into high, medium, or low categories according to their relation to the goals of the exhibit, as determined by the research team (see Table 1; note that the weighting is particular to this exhibit.). These relevance categories are assigned numerical weights in order to quantify and compare the substance of visitor talk across conditions. Using magnitude coding as a way of "quantizing" a phenomenon (Tashakkori & Teddie, 2010) permits the use of inferential statistics (Bernard, 2006; Saldaña, 2009) in order to compare the experimental conditions.

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

The codes assigned to the "low relevance" category – mostly INSTANTIATE and MANAGE subcodes – are all activities that are useful for grounding and coordinating the group learning experience and may serve as springboards for future dialogue, but are in and of 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 more directly make sense of the presented data by characterizing and contextualizing it (including instantiating outside knowledge to help with sense-making), clarifying the representational forms, and directing co-visitor's attention to an interesting element of the exhibit (which rises above a simple *instantiate* 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 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. Quantitative "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 number 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", 1) + *INSTANTIATES-geography* ("Brooklyn", 1) + EVALUATE-*characterizes* ("lots of", 2) + INTEGRATE-*connect:simple (dataset to geography)* (West Indians in Brooklyn", 1) + GENERATE-*confirms* ("oh yes … that is true", 3) = content score of 8. Whether this was 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 are intended to illuminate differences among 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. For example, consider again the 8-point statement above. An alternative method for evaluating visitors' dialogue might be to focus on the proportion of codes at each relevance level (high, mid, and low) applied in each session. If you look at proportions, that example becomes 3/5 (or 60%) low-relevance, and 20% mid-relevance and 20% high-relevance. So by that system, a much simpler statement like "It looks like there are more of them" is a 100% high-level statement, because it is an INTEGRATE-*compare* without any elaboration. But in terms of what it's adding to the conversation—keeping in mind sociocultural perspective on learning most studies of museum learning adopt—it is nowhere near the same level of substance. The simpler high-level statement 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. We do not assert there is an ideal ratio of low to mid to high statements. Five example sessions graphed in Figure 2 below demonstrate different profiles of productive conversation.

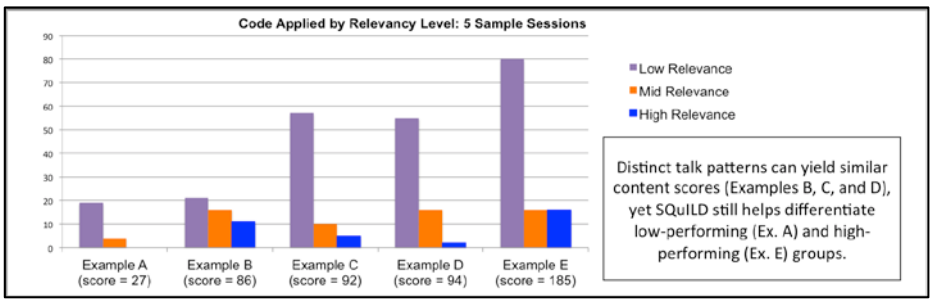

Figure 2. The SQuILD method is not intended to value a particular ratio as "best." Instead, dialogue is evaluated based on how well it aligns to learning goals and how many "hooks" visitors have to engage with each other and the presented content.

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 Figure 2 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 but receiving very little input from her companion. Examples B, C, and D, 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 others' 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 conducted by Roberts & Lyons (in progress) vetted this assumption by testing two alternative scoring proportions. 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 versions (Roberts & Lyons, in progress). 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, provide a valid quantitative measure for conducting A/B testing and informing exhibit design decisions.

Discussion

Understanding visitors' dialogue as they interact naturalistically with museum exhibits is of great interest to museum researchers but has traditionally been difficult to quantify for A/B testing. SQuILD accomplishes this through the combination several techniques. First, the context of the dialogue is maintained by coding all talk directly from the video recordings, which both illuminates the referents in visitors' conversations and allows segmentation of dialogue into a meaningful unit for the spontaneous, flowing discourse occurring in museums: the "idea unit." Because idea units as defined here, modified from Jacobs et al. (1997), can flow across multiple users as visitors co-construct the dialogue, they most accurately reflect the content of the shared meaning-making occurring in these interactions. Segmenting dialogue this way addresses a problem addressed by Allen (2002) in dealing with visitor discourse that tends to be "fragmented, ambiguous, or lacking clear referents" and that frequently involves repetition of words and phrases as members of a group echo each other. Allen dealt with this issue by coding only for the presence or absence of a type of talk during the entire interaction. Breaking the discourse into idea units that are then coded individually for presence of absence of a type of talk allows a clearer picture of the content of the dialogue to emerge.

By using simultaneous coding (Saldaña, 2009), a single idea unit can be coded with any number of codes, capturing not only the content of each statement but also the depth and complexity of the talk. Because each code applied, though productive, is not equally relevant to the learning goals of the exhibit, SQuILD utilizes magnitude coding (Saldaña, 2009; Miles & Huberman, 1994; Tashakkori & Teddie, 2010) to differentiate codes by their relevance to the exhibit's learning goals, much the way a teacher uses a rubric to quantify a student's piece of creative writing. By assigning weighted values to the codes applied, differences in the dialogue generated in each session are quantifiable and available for statistical comparisons.

We successfully employed this methodology in an *in situ* A/B comparison of different interaction designs for an collaborative data visualization exhibit (Roberts & Lyons, in progress), which allowed us to perform a number of statistical analyses exploring the relationship between exhibit use and the learning talk, and helped persuade stakeholders that a favorite design, despite being innovative and vetted in lab studies, was actually less successful for supporting learning talk. We expect that the SQuILD methodology will be useful to other researchers and designers developing open-ended computer-supported collaborative exhibits, in making similar assessments of the ecological validity of their design decisions on visitor learning. As museums embrace the dialogic model of education, they must concurrently embrace research methods suited to that model. The SQuILD methodology presented here takes steps toward that goal.

References

Allen, S. (2002). Looking for learning in visitor talk: A methodological exploration. Learning conversations in museums, 259-303.

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.

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.

Bernard, H.R.(2006). Research methods in anthropology:Qualitative & quantitative approaches. AltaMira Press.

Bitgood, S. (2000). The Role of Attention in Designing Effective Interpretive labels. *Journal of Interpretation Research*, *5*(2), 31–45.

Block, F., Hammerman, J., Horn, M., Spiegel, A., Christiansen, J., Phillips, B., ... & Shen, C. (2015, April). Fluid grouping: Quantifying group engagement around interactive tabletop exhibits in the wild. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 867-876). ACM.

Borun, M., Chambers, M., & Cleghorn, A. (1996). Families are learning in science museums. Curator, 39(2), 262-270.

Chi, M. T. H. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. Journal of the Learning Sciences, 6(3), 271–315.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. Journal for Research in Mathematics Education, 18, 382–393.

Davis, P., Horn, M. S., Schrementi, L., Block, F., Phillips, B., Evans, E. M., ... & Shen, C. (2013). Going deep: Supporting collaborative exploration of evolution in natural history museums. In Proceedings of 10th International Conference on Computer Supported Collaborative Learning.

Diamond, J., Luke, J. J., & Uttal, D. H. (2009). Practical Evaluation Guide: Tools for Museums & Other Informal Educational Settings (2 ed.). Lanham, MD: AltaMira Press.

Eberbach, C., & Crowley, K. (2005). From living to virtual: Learning from museum objects. Curator: The museum journal, 48(3), 317-338.

Falcão, T. P., & Price, S. (2009). What have you done! The role of 'interference' in tangible environments for supporting collaborative learning. In Proceedings of the Conference on Computer Supported Collaborative Learning. CSCL'09. ISLS, 324-334.

Falk, J. H., & Dierking, L. D. (2000). Learning from museums: Visitor experiences and the making of meaning. Altamira Press.

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 Math Education, 32(2), 124-158.

Heath, C., & vom Lehn, D. (2008). Configuring 'Interactivity' Enhancing Engagement in Science Centres and Museums. Social Studies of Science,38(1), 63-91.

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

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.

Leinhardt, G., & Crowley, K. (1998). Conversational elaboration as a process and an outcome of museum learning. Museum Learning Collaborative Technical Report (MLC-01). Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.

Leinhardt, G., & Knutson, K. (2004). Listening in on museum conversations. AltaMira Press.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook. Sage.

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1998). Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey. US Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328. Web site: http://nces. ed. gov..

Povis, K. T., & Crowley, K. (2015). Family Learning in Object-Based Museums: The Role of Joint Attention. Visitor Studies, 18(2), 168-182.

Roberts, L. C. (1997). From Knowledge to Narrative: Educators and the changing museum. Washington: Smithsonian Institution Press.

Roberts, J., and Lyons, L. (in progress). Bigger May Not Be Better: Impacts of Whole Body Interaction and Distributed Control on Visitors' Learning Talk at an Interactive Data Museum Exhibit.

Rounds, J. (2006). Doing identity work in museums. Curator: The Museum Journal, 49(2), 133-150.

Saldaña, J. (2009). The coding manual for qualitative researchers. Sage

Serrell, B. (1996). *Exhibit Labels: An Interpretive Approach*. Walnut Creek, CA: Altamira Press.

Tashakkori, A., & Teddlie, C. (2010). Sage handbook of mixed methods in social & behavioral research (2nd ed.). Los Angeles: SAGE Publications.

Vygotsky, L. S. (1978). Mind and society: The development of higher mental processes.Cambridge, MA: Harvard University Press.

Williams, A., Kabisch, E., & Dourish, P. (2005). From interaction to participation: Configuring space through embodied interaction. Proceedings of the Ubicomp 2005, LNCS 3660 (pp. 287-304).

Acknowledgments

This material is based upon work supported by the National Science Foundation under NSF INSPIRE 1248052.