Fluid Grouping: Quantifying Group Engagement around Interactive Tabletop Exhibits in the Wild

Florian Block[1], James Hammerman[2], Michael Horn[3], Amy Spiegel[4], Jonathan Christiansen[2*], Brenda Phillips[1], Judy Diamond[5], E. Margaret Evans[6], Chia Shen[1]**
[1] School of Engineering and Applied Sciences, Harvard University, [2] STEM Education Evaluation Center, TERC, [3]Northwestern University, [4]University of Nebraska-Lincoln, [5]University of Nebraska State Museum, [6]University of Michigan
[1]{fblock, cshen}@seas.harvard.edu

ABSTRACT
Interactive surfaces are increasingly common in museums and other informal learning environments where they are seen as a medium for promoting social engagement. However, despite their increasing prevalence, we know very little about factors that contribute to collaboration and learning around interactive surfaces. In this paper we present analyses of visitor engagement around several multi-touch tabletop science exhibits. Observations of 629 visitors were collected through two widely used techniques: video study and shadowing. We make four contributions: 1) we present an algorithm for identifying groups within a dynamic flow of visitors through an exhibit hall; 2) we present measures of group-level engagement along with methods for statistically analyzing these measures; 3) we assess the effect of observational techniques on visitors' engagement, demonstrating that consented video studies do not necessarily reflect visitor behavior in more naturalistic circumstances; and 4) we present an analysis showing that groups of two, groups with both children and adults, and groups that take turns spend longer at the exhibits and engage more with scientific concepts.

Author Keywords
Museums; learning; multi-touch tabletops; quantitative methods;

ACM Classification Keywords
H.5.3. Group and Organization Interfaces.

INTRODUCTION
One of the cornerstones of multi-touch technology is its ability to support simultaneous interaction between co-located users. In recent years, multi-touch research has moved out of its infancy and into "the wild" [13, 14, 15, 16, 18, 25, 27, 31]. Large multi-touch displays are now available through several commercial vendors and many real-world applications have emerged. Museums have received particular attention from the research community, as supporting meaningful social interaction is seen as central to learning in informal environments [12, 22]. Several research studies have established that multi-touch technology has the potential to engage visitors in fruitful collaborative learning [14, 15, 16, 25, 31].

Many of these studies are based on qualitative analysis. Surprisingly little quantitative evidence exists that explains clearly the factors contributing to visitor engagement and learning around interactive surfaces. There are numerous challenges and nuances present in assessing group interactions and engagement quantitatively. First, it is not clear what constitutes a visitor group. Museums can be crowded and chaotic environments where acquaintances and strangers form streams of ad hoc visitor groups around exhibits. The composition, size, and interactions within these groups change continuously and spontaneously over time. Studies of group engagement require a systematic definition of groups in these fluid settings as the behavior of individuals around an exhibit will be influenced by other people present. Second, study designs need careful consideration. Museums are free-choice environments. Common recruitment and observation techniques may affect this ecological property, have substantial effects on the very behavior under study and thus inadvertently bias the study outcome. Third, the type of application and its user interface can potentially influence visitor engagement.

In this paper, we present an empirical study of group engagement at the California Academy of Sciences in San Francisco, which receives 2 million annual visitors, and has a very diverse demographic audience. We used two different genres of interactive tabletop exhibits in order to identify factors that are consistent across both types of applications, and two types of study designs. Our concrete contributions are four-fold: after discussing related work and introducing our study designs we present an algorithm for identifying meaningful groups within a continuous flow of visitors.

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea.

http://dx.doi.org/10.1145/2702123.2702231

*Jonathan Christiansen is currently at Marywood University.
**Brenda Phillips is currently at Boston University.

Second we present refined measures of group-level engagement along with appropriate statistical analyses. We then make a methodological contribution by examining the effects of different data-collection techniques on group engagement. Finally, we assess the effect of group size, age composition, and overlap with other groups on visitor engagement with the exhibits and their scientific content.

RELATED WORK

Various factors influence group engagement around traditional exhibits. These include the adult and child gender [11, 21, 25], a group's size and age composition [4, 7, 8, 10, 20], and a group's strategy for learning [3]. However, it is unclear if any of the observed effects apply to interactive multi-touch exhibits, which may facilitate different forms of group engagement. Previous HCI research has studied group engagement around interactive tabletops in general [1, 13, 18, 27, 28, 29, 30, 31, 33, 35, 36, 37] and in museums in particular [14, 15, 16, 25, 31]. This includes understanding which gestures are used, how visitors approach surfaces in public spaces, transitions between user groups, and physical and verbal interaction between users [13, 14, 16, 18, 30].

While many "in the wild" studies have assessed engagement with multi-touch technology, the great majority primarily concentrate on *qualitative* analyses of observational data [1, 13, 14, 15, 16, 18, 25, 27, 31]. *Quantitative* analyses of group engagement are less common, particularly for studies that have been conducted outside of laboratory or controlled settings. Peltonen et al. measured distribution of group size and group overlap in front of a public interactive surface located in a shop window [30]; Horn et al. quantified holding times of recruited and non-recruited groups around an interactive game-based exhibit [15]; and Hinrichs et al. quantified occurrences of various multi-touch gestures around an interactive museum exhibit [14]. Other quantitative studies either focus on quantifying engagement around traditional, non-digital exhibits [9, 12, 20, 23] or are conducted in the lab with predetermined group sizes [7, 29, 33, 35, 36, 37].

Our aim, in contrast, is to add to this body of research studies by quantifying group engagement in an ecologically valid environment with natural visitor groups that dynamically form, reconfigure, and disperse. In this context, a significant challenge is to identify groups in a systematic way. Existing work has provided some basic mechanisms of identifying groups based on uninterrupted use [18, 27, 30]. This works for quiet venues, but is problematic when groups overlap [18]. Work on F-formations [19, 34] offers a way of defining and detecting free-standing groups through analyzing orientation and positioning of group members. In museums, however, such spatial characteristic are mostly predetermined by the design of the exhibit space [28]. Here we introduce an algorithm to identify groups based on their temporal formation within continuous periods of usage.

STUDY DESIGN

In this section we will first define measures of engagement for our quantitative study. We then describe our study design, data collection, and research questions.

Measures of Group Engagement

Several methods for quantifying engagement have been proposed. Dwell time (or holding time) is an established measure that is easy to capture and serves as a reasonable proxy for the depth of visitor engagement with exhibits [2, 3, 15, 17]. Following [13, 15, 23, 29], we also measure the frequency of physical and verbal behaviors that we detail in the Study Procedure section. Prior research offers some insight into which group factors may influence engagement and learning in museums. A series of studies have shown that *group size* influences engagement around exhibits [3, 6] and multi-touch tabletops [33]. Another set of studies provides evidence that *age composition* influences learning, highlighting that groups with both children and adults learn best [6, 9, 10, 11]. Third, there is qualitative evidence that the overlap between groups may influence engagement with interactive surfaces in naturalistic, walk-up-and-use scenarios [27, 30].

Applications

The type of application and its user interface can potentially influence visitor engagement around interactive tabletops. In this paper, we consider two different genres of tabletop exhibit to identify factors that are consistent across both applications. In this section we briefly describe both applications to give a sense of the different types of experiences that they provide.

The first application, called DeepTree [5], is an interactive visualization of the Tree of Life in which visitors can browse the evolutionary relationships of over 70,000 species. Visitors can "fly" through the tree to interesting species using a deep zoom interface, view descriptions and rich imagery about species they find, and learn about how any two species in the tree are related. Visitors experience DeepTree as an open ended, exploratory activity, in which they have free control over what they see and do. The second application, *Build-a-Tree* (*BAT*) [15], is a tree-building puzzle game in which visitors reconstruct the evolutionary relationships of different species in increasingly challenging levels. This is done by bumping species together in the correct order.

Both applications are the result of over two years of iterative testing and evaluation involving hundreds of users across several museums [5]. They are now on display in four major museums in the U.S. Both applications have been carefully designed to support collaboration and social engagement [4]. However, even though both applications involve evolutionary trees, the nature of the interaction is very different–one is an open-ended exploratory data visualization, and the other is a puzzle game.

Data Collection

There are several challenges to quantify group interaction in naturalistic settings. First, even though we measure engagement at a group level, we have to collect data at the subject level, as it is almost impossible for a real-time coder to reliably and consistently detect groups as they form. In fact, as we will outline below, even retrospectively identifying groups in a video recording is non-trivial. However, because subjects commonly overlap during the formation of groups, subject-level data is unlikely to be independent. This makes most traditional statistical analyses (such as ANOVAs) that assume independence of observations inapplicable on the subject level.

Secondly, the way in which we conduct observations may influence the way visitors interact with each other and the exhibit. Video recording in public spaces requires visitor consent in most countries. This, in turn, requires consent procedures that may introduce sampling bias and alter visitor behavior, thus undermining the ecological validity of the findings. Alternatively, engagement can be captured by manually coding visitor behavior in real-time. This allows for a more natural flow of visitors and is less intrusive. However, these types of observations present challenges for researchers to accurately measure the range of physical and verbal interactions between group members in real-time.

Our study design seeks to strike a balance between these two methods of observation. To assess the ecological validity of our results, we independently captured engagement twice. In one set of observations, we video recorded visitors using the DeepTree exhibit after obtaining informed consent (Video). For the second set of observations we coded visitor behavior in real-time at both the DeepTree exhibit and the Build-a-Tree exhibit. This was done without video recording or written consent procedures (Naturalistic). This resulted in three independent datasets (Set 1: *Video / DeepTree*; Set 2: *Naturalistic / DeepTree*, and Set 3: *Naturalistic / Build-a-Tree*). Across all three datasets, we used the same engagement variables and real-time coding scheme, as described in the next subsection. All the data were collected at the California Academy of Sciences in San Francisco during the same time of the year. IRB approval was obtained.

Study Procedure and Real-Time Coding

All observations were collected on the floor of the same museum with the general visitor population. In the *Video* condition, the area around the exhibit was cordoned off, and visitors signed a consent form before entering the area. Evaluation staff also actively recruited visitors who were nearby. After giving consent, participants were free to come and go as they pleased. Video and audio of their interaction at the exhibit were recorded. For the naturalistic condition, a sign next to the exhibit informed visitors of an ongoing study, and staff were available to answer questions. Otherwise visitors were entirely free to come and go.

For all of the datasets the same real-time coding scheme was used—the only difference was that coding in the *Video* study utilized the video replay, while in the *Naturalistic* study observers coded engagement on site. To make both study types comparable, the coding scheme was limited to what was accomplishable in real-time, even though the video would have allowed for more sophisticated coding.

The following events were captured. First, arrival and departure times were recorded, from which we could derive dwell times and overlap between visitors. Second, the age range of each visitor was estimated. Third, we developed and refined a coding scheme for social engagement. Our original scheme included 19 codes, but several of these occurred so rarely (<5% of the time) or so frequently (>95% of the time) that they presented insufficient variability and were eliminated in analysis. The following nine social engagement behaviors made up the final set:

- *Prevent Touch* (physical): One visitor prevents another visitor from touching the display.
- *Turn-Taking* (physical): Visitors take turns in taking control of the exhibit.
- *Two Manipulate* (physical): Two visitors manipulate the exhibit at the same time.
- *Pointing* (physical): A visitor points at an element on the screen but does not touch it.
- *Bio Question* (verbal)
- *Bio Statement* (verbal)
- *Negotiation* (verbal): Visitors negotiate what to do.
- *How-To Talk* (verbal): Visitors discuss how to operate the exhibit / user interface.
- *Enjoyment* (affective): A visitor expresses enjoyment.

Following standard museum practice [10, 11], interactions were coded in twenty (20) second intervals – noting that an identified behavior occurred during that interval. This means that even if an engagement behavior occurred multiple times during a twenty second interval, this behavior would be only recorded once during that interval. For each visitor, we can then calculate the percentage of time they experienced engagement behaviors by dividing the number of intervals in which they experienced each behavior by the total number of intervals they spent at the exhibit. For the remainder of this paper, we refer to these percentages as social engagement measures. Note that most social engagement measures can only occur when there are at least two subjects at the table. Consequently, visitors who spent most of their time interacting with the exhibit on their own were excluded from analysis of social engagement (but included in the analysis of dwell time).

Participants

Across all three studies, we collected data from 629 visitors over the course of 10 days, 169 for the *Video* study and 459 across the two *Naturalistic* studies, with 46 young children (~ < 5 years), 149 children (~ 5 – 12y) , 69 teens and 345 adults (20 unknown age group).

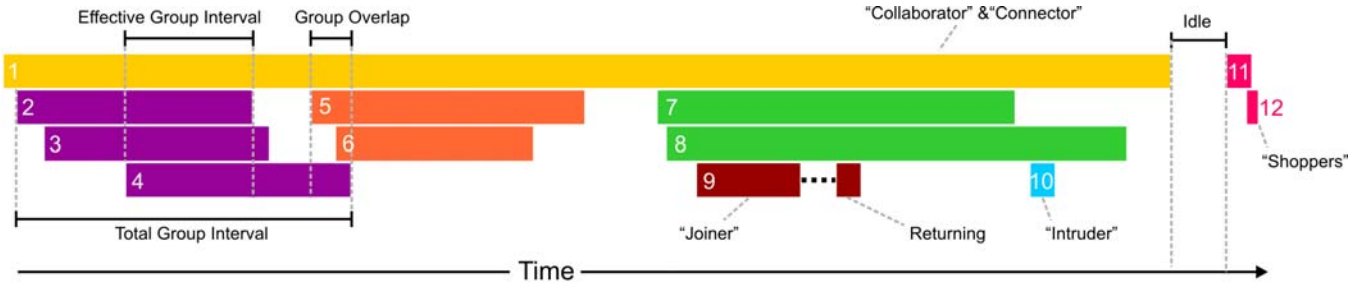

Figure 1. Example of a sequence of visitors spending time at the exhibit and clustering of exclusive groups (color coded).

Research Questions

Based on our measures of engagement, we focus on the following five research questions:

RQ1: How can visitor groups be characterized and quantified in a context in which groups spontaneously form, change over time, and disperse?

RQ2: How does the nature of the observational study affect visitor engagement with interactive exhibits?

RQ3: How does group size and age composition influence engagement?

RQ4: How do overlapping groups influence each other?

RQ5: Are certain types of social engagement associated with longer dwell times?

RQ1 contributes to an algorithmic definition of museum visitor groups. RQ2 addresses issues surrounding study methodologies. RQ3 through RQ5 intend to offer insights into group social engagement around science exhibits on a multi-touch table.

DEFINING GROUPS

While existing observational studies have looked at the formation of groups and their interaction in walk-up-and-use environments, we are not aware of systematic methods for identifying groups within a natural flow of visitors. Some studies have used idle times to cluster sessions and groups of continuous use [18, 27, 30]. However, this is not suitable for busy museums as groups may form and overlap *within* continuous periods of use. An example sequence of arrival and departures of several visitors is shown in Figure 1 (based on real patterns of visitor flow from our data). It illustrates that defining groups by idle time would give us only gross clusters that may contain multiple sub-groups that have come and gone, and thus would fail to detect more fine-grained group formations that overlap with one another. This makes quantitative analysis inaccurate and problematic, whether establishing factors such as group size and age composition, or measuring engagement or dwell time. For instance, it is not clear what we should consider to be the size, composition, and dwell time of this sequence of visitors, since there is only one visitor out of 10 who stays for the entire time, while the others spend vastly varying times at the exhibit in different group constellations.

Methodologically, the need for clearly defining groups arises from two requirements. First, we want to determine the representative group size and age composition in which the group members have *actually* spent most of their time together at the exhibit. Second, we need to analyze engagement on the group level for statistical validity. The experience of many visitors overlaps significantly, which introduces dependencies within our data that makes most statistical analyses inapplicable on the subject level. Through aggregating our engagement per group, we can avoid much of this dependence, which allows us to conduct sound statistical analysis on the group level.

Group Clustering Algorithm

For the purpose of this study, we define shared presence as temporal overlap expressed in percentage of time spent at the exhibit. For instance, if visitors overlap by only 5% of their combined time at the table, they are less likely to influence one another than visitors who spend all of their time together. Note that the percentage of overlap between two visitors is asymmetric when the dwell times of each visitor differ. For instance, visitor 2 in Figure 1 spends 100% of her time with visitor 1, but visitor 1 only spends a small fraction of her time with visitor 2. Our algorithm for identifying groups is based on groups of visitors who all *mutually* overlap for the majority of their time (>50%), so that for any two subjects A and B within a group, A spends most of her time with B, *and* B spends most of her time with A (e.g. visitors 2, 3, and 4). In other words, we select groups of visitors that *share most* of their experience – in this paper referred to as *exclusive groups*.

The core algorithm for this procedure works as follows: 1) calculate pair-wise percentage overlap between all subjects; 2) For each subject, check if there are existing groups in which the subject mutually overlaps with *all* members; 3) if there is exactly one group that meets this criteria, add the subject to the group; if there are multiple groups that meet this criteria, add the subject to the group in which the mutual overlap is highest; if there are no groups, create a new group with the subject as first member. We applied this algorithm to detect groups within our dataset, which clustered our 629 visitors into 354 groups across all three datasets. The visualized results form a sequence chart, such as shown in Figure 1. For illustration, Figure 1 uses color to code exclusive groups determined by the algorithm.

In a second pass, for each exclusive group, the algorithm finds non-members that overlap with *each* member by more than 50%. Note that this time, this overlap is not mutual as otherwise the visitor would have to be a member of the exclusive group. From the perspective of each group, we refer to such visitors who do not spend most of their time with a group, but with whom the group spends most of its time as *collaborators*. For instance, visitor 1, 7 and 8 in Figure 1 are collaborators from the perspective of visitor 9.

The distinction between exclusive group members and collaborators is important in two ways. First, if we want to determine the *actual* predominant configuration in which a group has interacted around the interactive tabletop, we have to consider both the exclusive members, as well as the collaborators of a group. For example, even though the size of the exclusive group of visitor 2, 3 and 4 is three, they have *actually* interacted in a group of size 4 (together with visitor 1). Second, if we aggregate engagement measures on the group level, we need to *exclude* collaborators as by definition, they spend less than 50% of their time with the group and their experiences are not representative of the group. For instance, if we calculate a representative dwell time for the exclusive group formed by visitors 2, 3, and 4, we should only average the dwell times of the exclusive members 2, 3 and 4, but not visitor 1 (who is a collaborator). Similarly, for social engagement measures, only the exclusive group members should be averaged, as their intervals of presence most accurately represent the frequencies of table-wide events of social engagement that happened within the duration of the exclusive group. Note that at the time of data collection, any of the Collaborators' actions would have affected the social engagement coding of other visitors at the table. Consequently, the exclusive groups' aggregated social engagement measures do implicitly reflect the social influence of Collaborators.

In the last step, the algorithm determines the group's effective size and composition (based on the ages of all members and collaborators), as well as its averaged dwell time and social engagement measures (based only on exclusive group membership).

Special Group Types

Based on a review of a visualization of all groups (as shown in Figure 1), we found several noteworthy patterns in the formation of exclusive groups. First, a few visitors (10 out of 629) spend extremely long periods of time (15+ minutes) at the exhibit, overlapping with multiple groups. Such visitors, who we refer to as *Connectors*, are identified as groups of size one by our algorithm (interacting alone) as there is no other visitor with whom they have spent more than 50% of their time. However, Connectors are different from other groups of size one as they do *not* spend the majority of their time alone, but within changing group configurations. As neither a clear group size nor age composition can be assigned for Connectors, we do not consider our 10 connectors as their own group for further analysis in this paper. Note that Connectors *are* counted as collaborators for any group they substantially overlap with, thus, influence the effective size and age composition of groups they collaborate with.

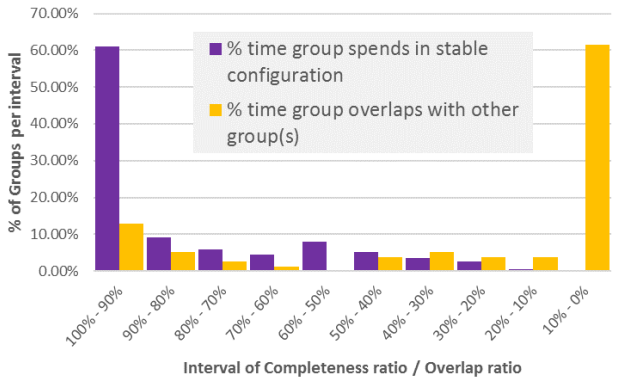

Figure 2. *Completeness Ratio* and *Overlap Ratio* distributions.

To have a better vocabulary for the qualitative discussion on group engagement, we also identified a few other patterns of visitor flow (who are included in the analysis): *Joiners* and *Intruders* are individuals or groups who approach the table when there is already at least one group present and leave before the initial group departs. However, *Joiners* end up staying with existing groups for more than 2 minutes, while *Intruders* leave before 2 minutes. *Shoppers* are individuals or groups who approach an empty table but leave before 2 minutes. The 2 minute threshold is based on our analysis of dwell time distribution as outlined in the next section.

Stability of Group Configurations and Group Overlap

For each exclusive group, we define the *Completeness Ratio* as the percentage of the total group time spent in the determined configuration of size and age composition. We calculate the Completeness Ratio by dividing the *effective group time* – in which *all* members *and all* collaborators were present – by the *total group time* – in which at least one of the exclusive group members was present (Figure 1 shows the example for exclusive group with members 2, 3 and 4). The higher the *Completeness Ratio*, the more representative the determined group size and age composition is with respect to the aggregated engagement measures. We also define the *Overlap Ratio* as the amount of time any visitor is present who is a member of another exclusive group and *not* a collaborator divided by the *total group time*. The lower the *Overlap Ratio*, the less potential there is for overlapping groups to affect social engagement measures and dwell times. Figure 2 shows a distribution of *Completeness Ratio* and *Overlap Ratio* across all of our groups. More than 60% of our groups spend more than 90% of their time in the determined configuration (purple bar on the left), and have less than 10% overlap with other groups (yellow bar on the right).

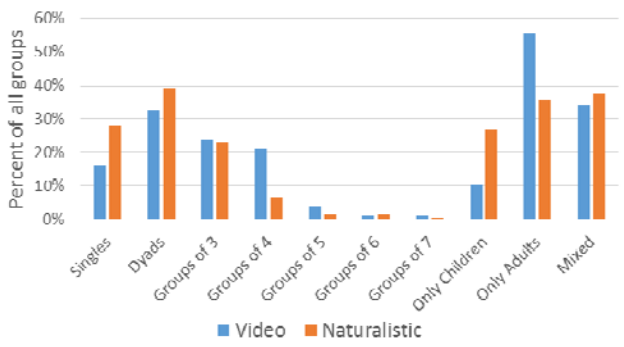

Figure 3. Frequency distribution of group size and age composition, by Study Type.

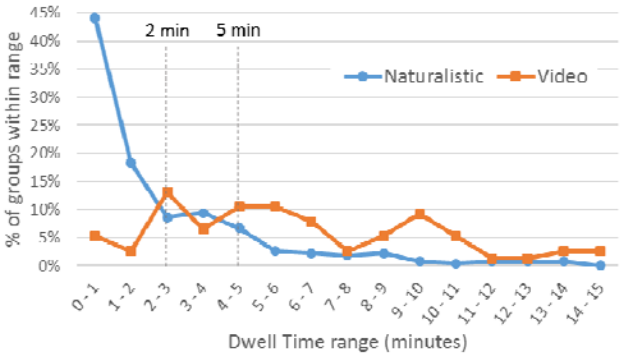

Figure 4. Distribution of dwell times per study type.

However, it is important to acknowledge the points between the extremes where around 40% of groups showed various degrees of completeness, as well as overlap with other groups. For instance, 12% of groups overlap with other groups by over 90% — this is the case for all Joiners and Intruders, who spend all their time at the table when an existing group is already present. This means that a minority of dependence of data remains. This can be used to quantify the general "messiness" of group formation in realistic flows of visitors. We could, of course eliminate all groups with overlap, or that are not in a stable configuration for most of their time. This, however, would also exclude cases that are characteristic of the museum context and important to consider when analyzing interaction around multi-touch surfaces. Consequently, we will include all cases in our analysis, and conduct statistical analyses that assume independence of data, even though we acknowledge that some degree of dependence remains.

COMPARING DIFFERENT STUDY TYPES

In this section, we address the question of whether the consent process and video recording in our *Video* setup significantly alter visitor group formation and engagement with the exhibit compared to the *Naturalistic* setups.

Frequencies of Group Size and Age Composition

We analyzed a total of 76 groups for our *Video* dataset and 267 groups for both of our *Naturalistic* datasets. Figure 3 shows the distribution of group size and age composition. The frequencies were significantly different for size and age categories shown in Figure 3 ($N = 343$, *Group Size*: $\chi^2(6) = 19.00$, $p = .004$; *Age Composition*: $\chi^2(2) = 12.32$, $p = .002$).

Across all datasets, 35% of groups were dyads (group size = 2), and 22% groups of three. Groups with sizes of 5 and over were rarely observed. In the *Video* data singles are significantly less common than in the *Naturalistic* data (16% vs. 28%), while groups of four are significantly more common in the *Video* data (21% vs. 7%). Overall, this shows that the *Video* study sample is biased towards higher group sizes. Further, *child only* groups were less frequent in the *Video* study (11% vs. 27%), while *adult only* groups were more common (55% vs. 36%). Mixed groups were similarly frequent at around 36% across both study types.

Distribution of Dwell Times

The distribution of dwell times is shown in Figure 4. Our two datasets only show comparable distributions of dwell times in the range between 2 and 5 minutes (*Video:* 25% vs. *Naturalistic:* 30%). The ranges below 2 minutes and above 5 minutes are inverse: around 60% of all groups in the *Naturalistic* data engage with the exhibit less than 2 minutes, while only 8% of groups in the *Video* study fall into this range. This shows that in a *Naturalistic* setting, the majority of groups are what we have labelled *Shoppers* and *Intruders*, and suggests that 2 minutes is a suitable threshold for their characterization. Inversely, while almost one half of all groups (49%) in the *Video* study stayed longer than 5 minutes, only 12% of groups stayed over 5 minutes in the *Naturalistic* study. Overall, this shows that dwell time in the *Video* data is strongly skewed towards longer durations compared to the *Naturalistic* data. A One-way ANOVA with factor *Study Type* shows that group dwell time is five times higher in the *Video* study (median 6.0 min) than in the *Naturalistic* study (median 1.2 min) ($F_{1,341} = 106.20$, $p < .001$). Note that Figure 4 clearly shows that dwell times are not normally distributed. Consequently, for all parametric tests, we use log-transformed dwell time, while reporting the median of the untransformed log times. Kolmogorov-Smirnoff tests showed that all log transformed data were normally distributed ($p > .200$ for all tests).

Differences in Social Engagement Behaviors

We ran Mann-Whitney Tests to compare social engagement measures across our *Video* and *Naturalistic* data (see Table 1 for all significant tests). Alpha was adjusted for multiple comparisons using Bonferroni ($\alpha = 0.005$). Groups in the *Video* study experienced approximately twice the amount of Turn-Taking and Pointing relative to the *Naturalistic* groups, while experiencing significantly less simultaneous interaction and How-To Talk. One interpretation of this result could be that groups who know that they are being recorded resort to more "orderly" forms of interaction, and are more hesitant to show confusion.

Measure	*U*	*P*	Video (median)	Naturalistic (median)
Turn-Taking	3325	< .001	20.1%	10.3%
Two Manipulate	4626	.003	31.1%	50.0%
Pointing	4164	< .001	20.5%	8.3%
How-To	4652	.003	23.4%	40.0%

Table 1. Differences in experienced engagement behaviors between *Video* and *Naturalistic* study (Mann-Whitney U).

Discussion

Even though our *Video* methodology encourages a natural flow of visitors and interaction around the table by allowing subjects who had given their consent to come and go as they pleased, there were a series of significant differences compared to the *Naturalistic* data. The consent procedure appears to have deterred / attracted some types of groups more than others, introducing a sampling bias towards larger groups and more adult-only groups and fewer child-only groups. After going through a consent procedure, visitors may also have been less likely to leave after only brief interaction, and more likely to spend time checking out the exhibit they "signed up for". Additionally, awareness of being part of a study may have encouraged a more thorough exploration of the exhibit and more orderly forms of social engagement.

CHARACTERIZING GROUP ENGAGEMENT

In this section, we will assess the effect of group size, age composition, and overlap with other groups on dwell time and social engagement. Given the significant differences between *Video* and *Naturalistic* data, we will analyze each dataset separately, instead of merging all data into a single analysis. Log-transformed dwell times were normally distributed for all analyses. Our social engagement measures are not normally distributed, even after log transformation. Consequently, all analyses of engagement measures use non-parametric Mann-Whitney tests. We adjusted alpha for multiple comparisons using Bonferroni.

Effect of Group Size and Age Composition

Naturalistic Data

We ran an ANOVA with three factors *Group Size* (*1, 2, 3*), *Age Composition* (*children only, adults only, mixed*) and *App Type* (*DeepTree, BAT*) with dependent variable dwell time. We included *App Type* to see if any effects are of a more general nature, or pertain to only one application type. Note that for this analysis we eliminated groups with sizes of four and larger, as we did not have sufficient data (only 7% of all groups consisted of four members, 2% of five, less than 2% of six, and less than 1% of seven members).

There was a significant main effect of *Group Size* ($F_{3,232} = 5.587$, $p = .004$) on *dwell time*. In the naturalistic studies, groups of two interacted approximately twice as long (median 1.9 minutes) as groups of other sizes ($p < .013$). There was no significant difference between single visitors (median 0.8 minutes), and groups of three (median 1.0 minutes) ($p = 1.000$). There was also a significant main effect of *Age Composition* ($F_{2,232} = 4.466$, $p = .013$) on *dwell time*. *Child only* groups (median 0.63 minutes) spent significantly less time at the exhibit than *Mixed* groups (median 1.6 minutes) ($p = .003$), and close to significantly less time at the exhibit than *adult only* groups (median 1.6 minutes) ($p = .091$). There was no significant difference between *mixed* groups and *adult only* groups in *dwell time* ($p = .954$). Our results show that groups do not stay longer than individuals, per se. Only dyads were found to engage significantly longer than visitors interacting alone.

There were no significant interactions between *App Type* and *Group Size* ($p = .946$), *App Type* and *Age Composition* ($p = .316$), or *Group Size and Age Composition* ($p = .388$), indicating that the measured main effects on dwell time were consistent across both applications, and independent from one another.

We also analyzed the effect of *Group Size* and *Age Composition* on social engagement. First, we compared differences in social engagement between groups of two and three (single visitors were omitted from the analysis of social engagement). Of our nine measures, only *Two Manipulate* differed significantly, with groups of three experiencing simultaneous interaction of two people 65% of the time (median), while groups of two experienced it only 39% of the time (median) ($U = 2075$, $p < .001$). While the rate of simultaneous interaction can be expected to go up with increasing group sizes, simultaneous interaction can also indicate increased conflict, which may be one reason for the lower dwell times for groups of three.

Social engagement was also significantly different for different age compositions. Table 2 shows that *mixed* groups experienced significantly more pointing, bio questions, bio statements, how-to talk and enjoyment than *children only* groups. Table 3 shows that compared to *adult only* groups, *mixed* groups experience significantly more preventing touches, two members interacting simultaneously, and negotiation. We believe this nicely captures the facilitating and moderating influence of adults on children in museums [6, 9, 10]. Note that for several social engagement measures, the majority of groups did not experience the relevant behavior at all.

Video Data

We conducted a similar ANOVA for the *Video* data (omitting the factor *App Type*, as we only have data for one application). However, none of the effects of *Group Size* or *Age Composition* from the *Naturalistic* data reoccurred in the *Video* data. There was *no* significant effect of *Group Size* ($F_{6,55} = .472$, $p = .627$) or *Age Composition* ($F_{2,55} = 1.022$, $p = .367$). There were also *no* significant differences of social engagement between groups of two and three, nor between *mixed* groups, *child only* groups, or *adult only* groups.

Measure	*U*	*p*	*Child only* (median)	*Mixed* (median)
Pointing	1472	< .001	0%	13.2%
Bio Question	1799	.001	0% (mean 6.5%)	0% (mean 15.9%)
Bio Statement	4164	< .001	0%	26.7%
How-To	1771	.003	0%	33.3%
Enjoyment	1833.5	.002	0%	4.3%

Table 2. Differences in social engagement in the *Naturalistic* data between *children only* groups and *mixed* groups (Mann-Whitney U).

Measure	*U*	*p*	*Mixed* (median)	*Adults only* (median)
Prevent Touch	2063.5	< .001	0% (mean 8.3%)	0% (mean 0.2%)
Two Manipulate	2074.5	< .001	37.5%	0%
Negotiation	1853.5	< .001	25.0%	0%

Table 3. Differences in social engagement in the *Naturalistic* data between *mixed* groups and *adult only* groups (Mann-Whitney U).

Measure	*DeepTree*	*BAT*
Turn-Taking	$r_s = .500, p < .001$	$r_s = .498, p < .001$
Pointing	$r_s = .613, p < .001$	$r_s = .639, p < .001$
Bio Questions	$r_s = .499, p < .001$	$r_s = .585, p < .001$
Bio Statements	$r_s = .599, p < .001$	$r_s = .586, p < .001$
Enjoyment	$r_s = .395, p < .001$	$r_s = .606, p < .001$

Table 4. Engagement measures that significantly correlate with dwell time across both applications in the *Naturalistic* study (Spearman's rho).

Effects of Overlap between Groups

We have determined that a substantial number of groups overlap. Based on our measure of overlap, we examine how the simultaneous presence of multiple groups affects dwell time and social engagement (RQ5).

Naturalistic Data

A Spearman correlation showed no significant correlation between *Overlap Ratio* and dwell time for the *Naturalistic* data ($r_s = .056$, $p = .363$). However, the more a group overlapped with another, the more frequently members experienced Two Manipulate, Pointing, Bio Statement, Negotiation, and How-To-Talk ($r_s > .198$, $p < .001$ for all correlations). This means that overlap was generally associated with elevated levels of social engagement around the table, but did not correlate with how long groups stayed at the exhibit. This is interesting, as we expected that social pressure would tend to make existing groups leave prematurely.

Video Data

For the *Video* data, there is a significant negative correlation between *Overlap Ratio* and dwell time ($r_s = -.440$, $p < .001$), meaning groups stayed longer the less they overlapped with other groups. This indicates that when part of our *Video* study, visitors may have more readily given up space for new groups approaching the table.

Correlation between Social Engagement and Dwell Time

Finally, to address RQ5 we ran non-parametric Spearman correlations between dwell time and each social engagement measure. We conducted this analysis separately for both applications in the *Naturalistic* dataset, focusing on significant correlations that we observed for both. We also ran all correlations for our *Video* data.

Naturalistic Data

Table 4 shows that across both applications, Turn-Taking, Pointing, Bio Questions, Bio Statements, and Enjoyment correlated significantly with dwell time. This means that the longer people interacted with the exhibit, the higher their rate of these engagement behaviors, including, importantly, enjoyment and engagement with scientific concepts.

It is important to note that correlations give us no information about the causal direction of the relationship. In other words, we cannot infer, for instance, that biological talk went up as a consequence of visitors staying longer, or if visitors stayed longer because they engaged in biological talk. Regardless, our findings provide support for the power of dwell time as a proxy for "good" interaction, as it correlates not only with orderly forms of physical activity, but also with enjoyment and engagement with scientific content. From a perspective of interaction design, it is interesting to note that while Turn-taking and Pointing clearly correlated with longer dwell times, Preventing Touch and Two Manipulate did not. This suggests that groups who take turns and do not interfere with one another also tend to spend longer at the exhibit and engage more frequently with the biological content. This is an interesting finding, as much of our interaction design for collaborative learning [4] has involved creating opportunities for simultaneous interaction.

Video Data

We ran the same set of correlations for the *Video* data and found that none of the social engagement measures correlated with dwell time ($p > .05$ for all).

DISCUSSION AND CONCLUSION

As demonstrated in this paper, a systematic and meaningful definition of what constitutes a group in a naturalistic flow of visitors is crucial for the quantitative analysis of engagement and interaction around interactive tabletops in museums. Our algorithm is based on a definition of shared experience as a metric for grouping. We chose this metric based on formal and informal observations of group engagement we conducted throughout the two year development process of both exhibits. We compared the outcome of the algorithm with manual grouping we had done based on the sequence charts. We found that, with a few minor exceptions, the algorithm presented in this paper concurred with our manual grouping, while identifying several mistakes and inconsistencies in the manual grouping. We hope that the proposed grouping algorithm as well as the analysis of overlap and group consistency will benefit future studies of group interactions in public spaces.

In the study setup that most closely reproduced naturalistic conditions, group size, age composition, and the occurrence of certain social behaviors significantly affect how groups engaged with our exhibits. All effects were measured independently across two different types of applications. Our findings have several implications for the design of learning experiences around interactive multi-touch exhibits and visitor research methodology:

Two is better than one and three. Our data suggest that groups do not necessarily engage longer than single visitors per se. We did find evidence that groups of two did engage with the exhibit for significantly longer than visitors who interacted alone, but groups of three did not spend significantly more time at the exhibit than visitors who were alone. This supports existing evidence that social dynamics beyond groups of two are more intricate [28, 33].

Design for small groups, particularly groups of two. Only a small number of people interact with the exhibit in groups of four or larger, so designing for larger groups may not be cost-effective. Instead, focus on designing interactive experiences around groups of two, as dyads spend the most time at the exhibit and show most engagement with the scientific content compared with groups of three.

Provide a meaningful single-user experience. Almost 30% of all visitors interacted with our exhibit on their own. Be sure that a design does not entirely rely on multi-user input and provide means for single visitors to have a meaningful experience.

Entwine playful elements and resources for advanced learners. Groups in which children and adults are mixed spend more time at the exhibit and verbally engage more with scientific content than children alone. Be sure to provide advanced scientific information (e.g. information overlays around the periphery of the display) that can give adults a more meaningful role as facilitators. This reinforces ideas supporting multi-level engagement [6] and synergistic scaffolding [24].

Many visitors will approach the exhibit while it is not in its initial state. Overlap between groups was common, which means that many visitors will enter the experience when others are already interacting. Traditional exhibit designs use reset mechanisms allowing newly joined visitors to start over. There might be opportunities for interactive digital exhibits to provide seamless experiences that are accessible regardless of the state of the exhibit.

Research methodology matters. Throughout the last two sections we have also established significant differences between our two study types. In our *Video* study, the effects on dwell time of group size, age composition, and social engagement were not significant, while more overlap between groups was associated with lower dwell times. These findings are exactly opposite of the *Naturalistic* datasets. Note that our data does not lend itself to identify exactly which factors underlying each study methodology contributed to these behavioral differences. However, we conclude that when quantifying engagement around exhibits, one should *attend carefully to the impact of consent procedures, as they can significantly affect group engagement*. Procedures that require formal consent to enter a small area cordoned off for videotaping significantly alter the flow of visitors to an exhibit and how they engage. If consent for video can be legally and ethically given at a distance in time and space (e.g., upon entering a museum or large exhibit space) then behavior and engagement may more closely resemble naturalistic conditions. If not, be cautious about using videotaped data to draw quantitative conclusions about dwell time or engagement.

LIMITATIONS

While we did take great care to highlight results that we think are ecologically most valid and that are of a more general nature, our results may be specific to the venue in which we have conducted the data collection. Our venue is a busy museum with 2 million annual visitors, and has a very diverse demographic audience. Some of our quantitative findings may or may not apply to other venues.

ACKNOWLEDGEMENTS

We would like to thank Anita Smith, Julie Shattuck, Lauren Hodge and Jim Galdos for their help in collecting the evaluation data, and the California Academy of Sciences for hosting this study. We would also like to thank Heather Lavigne, Joye Thaller and Anna Adams for their help with data processing and analysis. This work is partially supported by the National Science Foundation (DRL-1010889). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Antle, A.N., Bevans, A., Tanenbaum, J., Seaborn, K. & Wang, S. Futura: design for collaborative learning and game play on a multi-touch digital tabletop. In *Proc. TEI '11*, pp. 93-100, 2010.
2. Benton, D. P. Intergenerational interaction in museums. New York: Columbia University Teachers College. Ed. D. dissertation, 1979.
3. Bitgood, S., Kitazawa, C., Cavender, A. & Nettles, K.A. Study of Social Influence. Poster presented at the 1993 Visitor Studies Conference, Albuquerque, NM, 1977.
4. Block, F. Wigdor, D. Phillips, B. C., Horn, M.S. & Shen, C. FlowBlocks: a multi-touch UI for crowd interaction. In *Proc. UIST '12*, pp. 497-508, 2012.
5. Block, F., Horn, M.S. Phillips, B.C., Diamond, J., Evans, E.M., Chia Shen. The DeepTree Exhibit: Visualizing the Tree of Life to Facilitate Informal Learning. In *TVCG* 18(12), pp. 2789-2798, 2012.
6. Borun, M., Chambers, M. B., Dritsas, J. & Johnson, J. I. Enhancing Family Learning Through Exhibits. In *Curator* 40(4), pp. 279-295, 2010.

7. Buisine, S. Besacier, G., Aoussat, A. & Vernier, F. How do interactive tabletop systems influence collaboration? In *Computers in Human Behavior*, 28(1), 2012.
8. Cone, C., & Kendall, K. Space, time and family interaction: Visitor behavior at the Science Museum of Minnesota. In *Curator*, 21, 345-258, 1978.
9. Crowley, K., Callanan, M. A., Lipson, J. L., Galco, J., Topping, K. & Shrager, J. Shared scientific thinking in everyday parent-child activity. In *Science Education* 85(6), pp. 712-732, 2001.
10. Diamond, J. The behavior of family groups in science museums. In *Curator*, 29(20), pp. 139-154, 1986.
11. Diamond, J., Smith, A. & Bond, A. California Academy of Sciences discovery room. *Curator* 31, 157-166, 1988.
12. Falk, J. H., & Dierking, L. D. *Learning from museums: Visitor experiences and the making of meaning*. Altamira Press, 2000.
13. Fleck, R., Rogers, Y., Yuill, N., Marshall, P., Carr, A., Rick, J. & Bonnett, V. Actions speak loudly with words: unpacking collaboration around the table. In *Proc/ ITS'09*, pp. 189-196, 2009.
14. Hinrichs, U., & Carpendale, S. Gestures in the wild: studying multi-touch gesture sequences on interactive tabletop exhibits. In *Proc. CHI'11*, 3023-3032, 2011.
15. Horn, M., Leong, Z., Block, F., Diamond J., Evans, E.M., Phillips, B. & Shen C. BATs and APEs: Designing an interactive tabletop game for natural history museums. In *Proc. CHI'12*, 2059–2068, 2012.
16. Hornecker, E., "'I don't understand it either, but it is cool" - visitor interactions with a multi-touch table in a museum," In *Proc. ITS'08*, pp.113-120, 1-3 Oct. 2008.
17. Humphrey, T. and Gutwill, J. Fostering active prolonged engagement: The art of creating APE exhibits. Exploratorium (2005).
18. Jacucci, G., Morrison, A., Richard, G.T., Kleimola, J., Peltonen, P., Parisi, L. & Laitinen, T. Worlds of information: designing for engagement at a public multi-touch display. In *Proc. CHI'10*, pp. 2267-2276, 2010.
19. Kendon, A. The F-Formation System: Spatial-Orientational Relations in Face to Face Interaction. In Man Environment Systems(6), pp. 291-296, 1976.
20. Koran, J., Koran, M. & Longino, S. The relationship of age, sex attention, and holding power with two types of science exhibits. In *Curator*, 29(3), pp. 227-224, 1986.
21. Kremer, K. & Mullins, G. Children's gender behavior at science museum exhibits. *Curator*, 35(1), 39-48, 1992.
22. Leinhardt, G., Crowley, K., & Knutson, K. (Eds.). *Learning conversations in museums*. 2002.
23. Leinhardt, G., & Crowley, K. Museum learning as conversational elaboration: a proposal to capture, code, and analyze talk in museums. *Museum Learning Collaborative,* 1998. *http://museumlearning. org/paperresearch. html (accessed March 1, 2007)*.
24. Lyons, L. et al. Synergistic Scaffolding of Technologically-enhanced STEM Learning in Informal Institutions. In *Proc. ICLS'14 Vol. 3*, 1456-1465, 2014.
25. Ma, J., Liao, I., Ma, K., Frazier, J. Living Liquid: Design and evaluation of an exploratory visualization tool for museum visitors. In *TVCG* 18(12), pp. 2799-2808, 2012.
26. McManus, P. It's the company you keep… The social determination of learning-related behavior in a science museum. In *International Journal of Museum Management and Curatorhip*, 53, pp. 32-50, 1987.
27. Marshall, P., Morris, R., Rogers, Y., Kreitmayer, S. & Davies, M. 2011. Rethinking 'multi-user': an in-the-wild study of how groups approach a walk-up-and-use tabletop interface. In *CHI'11*, pp. 3033-3042, 2011.
28. Marshall, P., Rogers, Y. & Pantidi, N. Using F-formations to analyse spatial patterns of interaction in physical environments. In *CSCW'11*, 445-454, 2011.
29. Olson, I.C., Leong, Z.A., Wilensky, U. & Horn, M.S. It's just a toolbar!: using tangibles to help children manage conflict around a multi-touch tabletop. In *Proc. TEI'11*, pp. 29-36, 2010.
30. Peltonen, P., Kurvinen, E., Salovaara, A., Jacucci, G., Ilmonen, T., Evans, J, Oulasvirta, A. & Saarikko, P. It's Mine, Don't Touch!: interactions at a large multi-touch display in a city centre. In *CHI'08*, 1285-1294, 2008.
31. Rick, J., Marshall, P., & Yuill, N. Beyond one-size-fits-all: how interactive tabletops support collaborative learning. In *Proc. IDC'11*, pp. 109-117, 2011.
32. Rosenfeld, S. & Turkel, A. A naturalistic study of visitors at an interactive mini-zoo. *Curator*, 25(3), 187-212, 1982.
33. Ryall, K., Forlines, C., Shen, C. & Morris, M.R. Exploring the effects of group size and table size on interactions with tabletop shared-display groupware. In *Proc. CSCW'04*, pp. 284-293, 2004.
34. Setti, F., Russell, C., Bassetti, C. and Cristani, M. F-formation Detection: Individuating Free-standing Conversational Groups in Images. Preprint on arXiv.org, http://arxiv.org/pdf/1409.2702.pdf, 2014.
35. Shaer, O., Strait, M, Valdes, C., Feng, T., Lintz, M. & Wang, H. Enhancing genomic learning through tabletop interaction. In *Proc. CHI'11*, pp. 2817-2826, 2011.
36. Tang, A., Tory, M., Po, B., Neumann, P. & Carpendale, S. Collaborative coupling over tabletop displays. In *Proc. CHI'06*, pp. 1181-1190, 2006.
37. Tse, E., Shen, C., Greenberg, S., Forlines, C. How Pairs Interact Over a Multimodal Digital Table. In *CHI'07*.