

Towards Disrupting the Status Quo of Exhibit Design:  
A Data Reanalysis to Support Future Research

Toni Dancstep (née Dancu) and Joshua P. Gutwill  
Exploratorium, San Francisco, California

2019

Note:

This unpublished manuscript is aimed solely at researchers to inform future studies rather than at practitioners to improve exhibits. Although we typically try to reach both audiences in our work, we encountered an unexpected issue while working on the dissemination of this reanalysis: We realized that the results in this paper could overpower the original results in the Exhibit Designs for Girls' Engagement study for several reasons. First, when faced with the option to design for all visitors or design for girls, many practitioners expressed that they would choose the former even though the latter is necessary given current disparities in the field. Second, the reanalysis is descriptive while the original EDGE analysis was parametric; unfortunately, that subtle difference was difficult for our readers and reviewers to keep in mind when thinking about such a broad set of data. By focusing on researchers as the audience for this article and describing the original EDGE study at some length, we hope to mitigate the possibility that this reanalysis will be applied to practice or overshadow the original EDGE work.

### **Abstract**

This article describes the reanalysis of a large dataset containing engagement information for 60 exhibit design attributes at 301 exhibits across three institutions. The reanalysis revealed relationships between existing science museum exhibit designs and four measures of children's engagement (use, time spent, return, and behavior), providing baseline data across exhibits and engagement results for all design attributes. We caution against using the results to inform practice as research on inclusion has found that current exhibit designs may promote disparities among learners. Instead, we present these results to learning science researchers in hopes they will identify promising studies that can inform the field. In our own exploration, we categorized design attributes as (a) positively associated with engagement here and in prior studies, (b) common but understudied, (c) rare and promising, and (d) possibly inhibiting engagement, making suggestions about how future research can disrupt the status quo to create more welcoming learning environments.

*Keywords:* learning research; informal learning, learning sciences; exhibit design attributes; engagement; STEM

## **Toward Disrupting the Status Quo of Exhibit Design: A Data Reanalysis to Support Future Research**

For decades, the museum field has sought out exhibit designs that will enhance visitor engagement. Faced with myriad design approaches and the expense of conducting research in ecologically valid museum settings, learning researchers in informal environments typically must focus on only the few designs they believe are most important. Here, we describe the reanalysis of a sizable dataset in the hopes that our colleagues may use it to inform research regarding exhibit designs to advance free-choice science education.

Recently, the first author undertook a comprehensive study of exhibits in order to identify designs that most successfully engage girls ages 8-13 years. That project, called Exhibit Designs for Girls Engagement (EDGE), explored nearly 60 exhibit design attributes and found nine that consistently and significantly corresponded with positive engagement for girls in science museums (Dancstep & Sindorf, 2016, 2018a, 2018b). In this article, we briefly describe the methods and results of EDGE to set the scene for a reanalysis of that project's data. The reanalysis has a different goal from the original study, and combines data gathered on both girls and boys.

Our aim here is to provide a snapshot of exhibit use and design in order to answer new questions using the EDGE dataset:

1. What does typical child engagement look like at science exhibits in our sample? That is, what are the baseline levels of exhibit engagement for children (ages 8-13)?

2. What is the relationship between each of the design attributes and children's engagement? How can researchers employ those relationships to identify promising areas of study?
3. How common or rare are those design attributes in science museum exhibits? Can this information also aid researchers in focusing on productive areas of study?

The reanalysis offered in this article describes the status quo in science museum exhibit design, a status quo that may contribute to inequities, such as imbalances in science center engagement in favor of white, middle-class, and male visitors (Borun, 1999; Dawson, 2014a, 2014b; Diamond, 1994; Feinstein, 2017; Girls Inc., 2004; Hamilton, Michael, Kupermintz, Kerkhoven, & Snow, 1995; National Science Foundation, 2003, 2007). Applying the results of this reanalysis to practice could inadvertently strengthen current disparities. Consequently, our reanalysis does not prescribe particular exhibit designs, but by significantly expanding on prior research, may lend additional evidence in support of certain designs and be used to identify other designs that merit additional investigation.

### **Literature Review**

Identifying effective exhibit designs has been a goal of various museums and educators for over twenty-five years (e.g., Ansbacher, 1999; Bailey, Bronnenkant, Kelley, & Hein, 1998; Borun et al., 1998; Dancstep, Gutwill, & Sindorf, 2015; Falk, 1997; Garibay, 2013; Humphrey & Gutwill, 2005; Perry, 1993, 2012), and the field has begun to build a set of design principles or attributes that generally seem to promote learning. To lay the foundation for the reanalysis of the EDGE project's dataset, we

review prior research on exhibit designs that promote visitor engagement and learning.

Early studies on exhibit designs focused on the text and layout of labels for object-based exhibits. For example, investigators found that headers and chunked text help visitors scan a label for the information they need (Bitgood & Gregg, 1986; Bitgood & Patterson, 1993; Screven, 1995; Serrell, 1996, 2015), and shorter labels are read more thoroughly than longer labels (Bitgood & Gregg, 1986). Surprisingly, shorter labels even entice visitors to read a greater number of words than longer labels (Bitgood & Thompson, 1989). These early studies showed that general design principles could be applied to labels to improve museum visitors' learning experiences with object-based exhibits.

During that same period of time, Deborah Perry published her influential dissertation investigating the impact of exhibit design on learning and engagement (Perry, 1989). Perry (2012) recently expanded this work into a book called, *What Makes Learning Fun? Principles for the Design of Intrinsically Motivating Museum Exhibits*. By analyzing visitors at the Colored Shadows exhibit at the Indianapolis Children's Museum, Perry developed an instructional-design theory with six overarching design principles for meeting visitors' learning needs:

communication, curiosity, confidence, challenge, control, and play. Within each of the six principles, Perry proposed specific design prescriptions, such as:

- Direct visitors' attention to the important parts of the exhibit, object, or artifact.
- Provide a variety of different types of activities.

- Design spaces that encourage members of visiting social groups to stay together and in close proximity.

Perry's total number of design prescriptions is 65, each of which supports her model of visitor learning that includes motivations, engagements, and outcomes. This research provided a framework for developing exhibits based on learners' motivations and needs, helping to launch the field into an examination of exhibit designs.

The landmark PISEC study (1999) followed, finding that four interactive exhibits were more effective at promoting family learning when they had seven key design attributes: multisided, multiuser, multi-outcome, multimodal, accessible, readable, and relevant. Many of Perry's design prescriptions fit with the PISEC design attributes. These publications inspired several exhibitions and research studies across the science museum field. For example, the Museum of Science in Boston developed multiuser, multi-outcome exhibits in its Investigate! exhibition (Bailey et al., 1998) and the Science Museum of Minnesota elevated the attribute of multi-outcome even further in its Experiment Benches exhibition (Perry, 1993). The Exploratorium fleshed out some of the PISEC designs with a research and development project called Active Prolonged Engagement (APE), which found that exhibits engaged visitors more deeply when they offered multiple related options, allowed for multiple users who could not interfere with each other, and provided forking paths of multiple intriguing outcomes (Humphrey & Gutwill, 2005).

More recently, the Design Zone project at OMSI and the Geometry Playground project at the Exploratorium investigated the impact of APE and other design attributes

on mathematical thinking among visitors. Focusing on algebraic thinking, the Design Zone team explored design attributes such as multiuser interfaces, multi-option math challenges, cooperative experiences, and competitive activities. Ultimately, the exhibition demonstrably encouraged prolonged engagement, inspired large percentages of visitors to pose challenges for themselves, and sparked reasoning about functional relationships (Garibay, 2013). Working at the Exploratorium in the domain of geometry, we concentrated on spatial reasoning (Dancstep et al., 2015). Although the geometry team also utilized APE designs, they explored the impact of exhibit scale, comparing visitors' experiences at tabletop exhibits to those at whole-body, immersive exhibits. (The exhibits differed in size, but shared identical phenomenological content and similar visitor activity.) A quasi-experimental study found that tabletop designs promoted deeper spatial reasoning, while immersive designs triggered more return visits and were described more positively by visitors.

Each of these research or development projects involved a relatively small number of exhibits and design attributes. Many of the label studies concentrated on only single exhibits or exhibitions.<sup>1</sup> Although Perry's study involved recording 611 visitor groups, those groups interacted with only two versions of the Colored Shadows exhibit (380 at original and 231 at revised). The PISEC study looked at the largest number of design attributes to date, seven; each of four exhibits was studied in both its original condition and an enhanced condition instantiated with the seven attributes. The total number of participants involved was 400, with 50 in each condition of each exhibit.

Studies of exhibit designs at Museum of Science in Boston, OMSI, Ruben Fleet Science

---

<sup>1</sup> Although some museums use the word exhibit to refer to an entire collection, we refer to the individual elements or components of an exhibition as exhibits.

Center, Exploratorium, and Science Museum of Minnesota have ranged from single exhibits to exhibition collections of 30–60 exhibits, and typically focus on only a few design attributes (Bailey et al., 1998; Dancstep et al., 2015; Garibay, 2013; Humphrey & Gutwill, 2005; Perry, 2012; Sandifer, 2003). The relatively small number of exhibits and attributes in these studies may have limited what we are able to learn about those attributes, how they compare to one another, and how they hold up across a large number of exhibits and visitors.

As mentioned at the top of this article, the first author conducted the large-scale EDGE study to investigate the relationship between female-responsive exhibit designs and girls' engagement. Across three science museums, researchers followed 456 girls and 450 boys as they moved among 301 STEM exhibits that represented nearly 60 distinct design attributes. EDGE was motivated by research indicating that girls visit science museums less frequently than boys (Borun, 1999; Hamilton et al., 1995; Lee & Burkam, 1996; National Science Foundation, 2003), and that once inside, girls often have less engaging experiences at exhibits than boys (Diamond, 1994; Girls Inc., 2004; Greenfield, 1995; Kremer & Mullins, 1992; National Science Foundation, 2003; Verheyden, 2003). Ultimately, the project identified nine design attributes that were most important for engaging girls at STEM exhibits, none of which were harmful to boys' engagement. (For a list, see Dancstep & Sindorf (2016, 2018b)).

This article introduces a new analysis of the EDGE data, one focusing on the combined engagement of girls and boys at the 301 exhibits. We begin with a brief description of the methods employed in this reanalysis. Importantly, EDGE advocated for *changing* science museum practice by making exhibit design approaches more inclusive.

Our reanalysis provides a snapshot of the status quo, and identifies fruitful areas for future research. Consequently, the analytic methods differ somewhat.

### **Data Collection Methods**

To gather the original data for this reanalysis, our research followed a four-step process:

1. Identify design attributes that had the potential to engage females.
2. Assess exhibits across three museums for those attributes.
3. Measure engagement of girls and boys across those exhibits.
4. Find the design attributes that consistently, positively, and strongly related to girls' engagement. (This step differs in the present reanalysis.)

What follows is a brief description of each step.

#### **1. Identify Potential Design Attributes**

Through an extensive review of prior research from various fields; input from experts in design, gender, equity, and museum learning; and discussions with advisory groups of girls and exhibit practitioners, the first author's team compiled a list of nearly 60 specific design attributes (DAs) that had the potential to engage girls. The list of DAs covered a wide range of topics, such as seating, size of exhibit, number of users, look and feel, type of interaction, and label characteristics, just to name a few (for the list of DAs see Table V).

#### **2. Assess Design Attributes at 301 Exhibits**

The research included 301 exhibits at three U.S. institutions that varied in size, region, demographics, and approach to exhibitry. At a large West Coast science center, a

large Midwest science center, and a medium-sized Southwest science center, exhibits on the topics of physics, engineering, mathematics, and perception were analyzed for the presence or absence of each design attribute.<sup>2</sup> Therefore, the results are limited to particular topic areas, and any summary of prevalence of the design attributes may not generalize beyond our (nonrandom) sample of exhibits and institutions.

Multiple researchers examined each exhibit for each of the 60 design attributes. For the majority of attributes, researchers noted their presence or absence. A few attributes required a rating on a scale, signaling the extent to which the exhibit design embodied that attribute (e.g., how *whimsical* or *homey* the design appeared); in those cases, the attributes were later categorized into presence or absence based on the data. All items were examined and adjusted for face and content validity via discussions with advisory groups and practitioners.

During data collection on the exhibit designs, the team utilized a process of double coding and then resolving disagreements. Two researchers independently scored all exhibits for each design attribute. The pairs of researchers then compared their results, discussing and resolving on final codes for each attribute at each exhibit. This process of double coding followed by resolution of disagreements was used to improve reliability (Baxter & Jack, 2008; Olszewski, Macey, & Lindstrom, 2006). For those interested, the protocol and coding scheme can be found on our website (Sindorf & Danstep, 2015).

### **3. Measure Engagement**

The project employed traditional timing and tracking methodology (Serrell, 1998, 2010; Yalowitz & Bronnenkant, 2009). This method constitutes unobtrusive observation

---

<sup>2</sup> Size of museum based on operating revenue (Association of Science-Technology Centers, 2016).

of participants as they move through a museum space, and is commonly used for understanding visitors' engagement and behaviors in museum settings (Serrell, 1998, 2010). There are many ways to measure exhibit engagement, and each offers important information about different aspects of the exhibit's design and its visitor experience (Sanford, 2010). Pairs of researchers, blind to the study's purpose, simultaneously tracked 15% of participants to check inter-rater agreement. We explored four measures of engagement at all 301 exhibits.

**Exhibit Use.** Do children stop to use the exhibit? The attraction power of an exhibit in free-choice settings is often measured by the percentage of visitors who stop to use an exhibit (Diamond, 1986; Falk, 1983; Humphrey & Gutwill, 2005; Sandifer, 2003; Serrell, 1998). In our dataset, Exhibit Use was calculated by the percentage of children in the study who touched an exhibit for three or more seconds; or stopped, feet planted, facing the exhibit for five or more seconds. Inter-rater agreement, collected for 15% of the participants, was high at 93% for all instances of Use.

**Return Visits.** Do children go back to use the exhibit again? Return visits may be an important indicator of engagement in a free-choice environment where learners can vote with their feet (Diamond, 1986; Falk, 1983; Sandifer, 2003; Serrell, 1998). Return Visits was computed as the percentage of visitors who used an exhibit and who, after leaving, chose to return to use it again. Participants were considered to have returned to an exhibit if they met the requirements for Use at two or more unique times during their visit. Inter-rater agreement across all instances of Use, including Return Visits, was 93%.

**Time Spent.** How much time do children spend at the exhibit? Holding power is typically assessed by the average time visitors spend at an exhibit (Diamond, 1986; Falk,

1983; Humphrey & Gutwill, 2005; Sandifer, 2003; Serrell, 1998). Time Spent adds together the durations of all instances of Use for an individual, from when the child first uses an exhibit to when the child walks away, *or* faces away from the exhibit for five or more seconds. To measure Time Spent at an exhibit, the average time is calculated for all children at that exhibit. Inter-rater agreement for Time Spent across all instances of exhibit Use identified by either rater was high with Spearman's Rho at .94 (Ferguson, 2009).

**High Engagement Level.** Do children repeat, vary, and/or talk about aspects of the experience while at the exhibit? One of the most powerful approaches for assessing engagement involves directly observing visitors' various learning behaviors at exhibits (Barriault & Pearson, 2010; Borun & Dritsas, 1997; Humphrey & Gutwill, 2005; Sanford, 2010; Van Schijndel, Franse, & Raijmakers, 2010). For this measure, we drew on two existing metrics (Barriault & Pearson, 2010; Van Schijndel et al., 2010) to create a scale that would provide a sense of the depth of children's engagement at exhibits. Using the Engagement Level Scale (Table 1), researchers recorded the highest level of engagement each participant reached at each exhibit, from inattentive engagement to actively repeating, varying, and/or sharing their explorations. Collapsing across categories, the analysis used the percentage of children who reached the third level of engagement: High Engagement Level.<sup>3</sup> Inter-rater agreement, collected for 15% of the participants, was high at 92%.

---

<sup>3</sup> The majority of children reached Level 3, so we decided that the most useful unit of analysis would be whether or not participants reached the High Engagement Level.

Table 1

*Engagement Level Scale*

Level	Description	Examples (may do one or more of these)
1	Child passively or inattentively engages with the exhibit.	Observes inattentively, gaze may wander May passively touch/lean on exhibit May seem bored, inattentive
2	Child actively and attentively engages with the exhibit and may seem to enjoy the experience.	May watch attentively May touch exhibit elements Focuses on exhibit May show strong emotions; signs of interest/enjoyment/frustration
3	Child explores the parameters of the exhibit and may want to share the experience with others.	Actively and attentively applies repetition or variation to activity (tries multiple things, tests variables) Shares exhibit experience with others (on-topic talk, collaboration, showing others, calling others over, etc.)

*Note.* This table was originally presented in Dancstep and Sindorf (2018b).

To gather engagement data, we randomly selected children ages 8–13 attending a museum with their families and invited them to participate in the study. We tracked 906 children as they used or passed by the exhibits in a contained area of the museum in a free-choice manner. Specifically, the researchers tracked 637 children across 213 exhibits at the large West Coast science center; 131 children at the 70 exhibits at the large Midwest science center; and 138 children at the 18 exhibits at the medium Southwest science center. For some measures, it was necessary to remove a few exhibits from the analysis.<sup>4</sup>

<sup>4</sup> When relating each design attribute to the engagement measures, we removed confounding exhibits. That is, to minimize the confounding effects of exhibit location, we removed any exhibits that were in low-traffic areas when looking at the correlation between Use and the DAs. In the end we had a final sample size of 269 exhibits for the analyses exploring the relationships between the DAs and Use. Similarly, for some exhibits, Time Spent is not a good indicator of engagement. We identified and removed exhibits that were designed to offer a very short experience (Humphrey & Gutwill, 2005) when correlating Time Spent with the DAs. The final exhibit sample size for the analyses representing the relationships between DAs and Time Spent was 279 exhibits. All exhibits were included in the baseline analyses, which aim to provide a broader picture of exhibit engagement.

Prior research has found that visitors use only a portion of exhibits in any given museum space (Serrell, 1998), and the design of the space affects exhibit use and engagement (Adamson, 2008; Allen, 2004; Falk & Dierking, 2000). The layout and exhibit density of the three museum spaces studied in EDGE differed, contributing to the external validity and generalizability of the results. The West Coast museum offered its 213 exhibits in a large, open area of approximately 24,000 square feet (8.9 exhibits per 1000 sq ft), with several corridors leading in and out, a single wide path for foot traffic through the center, and few walls or smaller rooms. The Midwest museum presented 70 exhibits in a total exhibit area of about 15,000 square feet (4.7 exhibits per 1000 sq ft), and the spaces were broken up by empty regions or low walls as well as a few small adjoining rooms. The exhibits were on the bottom floor of the institution; three other exhibition spaces on that level were not included in the study. The Southwest museum space contained all 18 exhibits in a single small room of approximately 2,800 square feet (6.0 exhibits per 1000 sq ft) on the second floor of the institution. Figure 1 shows the rough layout of exhibit spaces within each institution, approximately scaled by size of the space.

The West Coast museum offered the greatest number of exhibits across the largest area for the study, so we recruited a higher number of participants there. We used analytic techniques to deal with the different sample sizes in the EDGE analyses but were not able to address this disparity in the reanalysis (see next section). To obtain sampling parity by gender, age, and interior location, we oversampled within each institution when any imbalances greater than five occurred between girls and boys, younger and older children, or recruit locations on a museum's floor. See Table 2 for participants'

demographics. Although the survey offered a non-binary and open-ended gender identity option, no children chose it. In addition, the sample contained too few children of color for meaningful analyses by ethnicity.

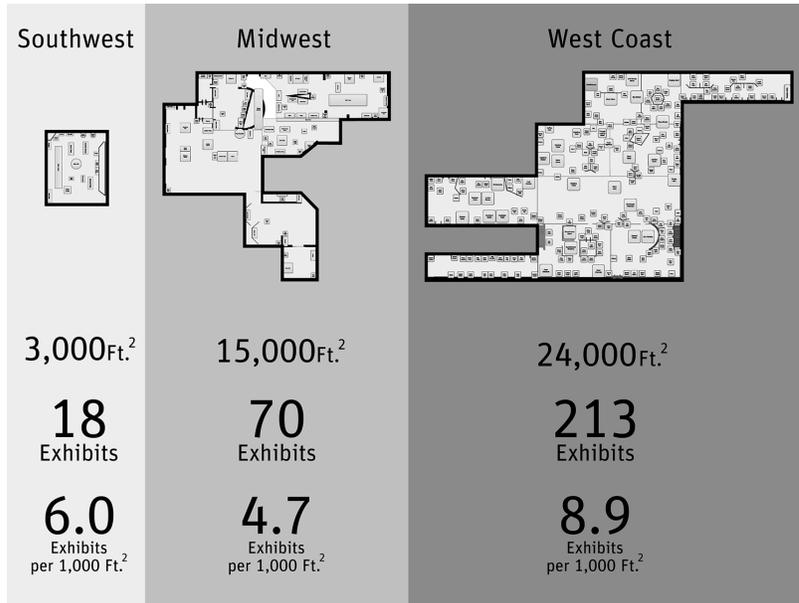


Figure 1. Exhibits and spaces at the three participating institutions.

Table 2

*Participant Demographics*

Demographic Information	n	Percentage
Total participants	906	100%
Gender		
Male	450	50%
Female	456	50%
Another category	0	0%
Race/Ethnicity		
Caucasian or White	606	67%
Hispanic or Latino	103	11%
Asian	91	10%
African-American or Black	43	5%
American Indian and Alaska Native	16	2%
Native Hawaiian and Other Pacific Islander	14	2%
Other	10	1%
Unspecified	78	9%
Attendance		
First time	729	80%
Return	154	17%
Missing data	23	3%

*Note.* The survey offered three options for gender, including *Another Category* with a write-in field. Ethnicity totals for each category include multi-racial/multi-ethnic participants; thus totals per category may be greater than participant totals. This table originally presented in (Dancstep & Sindorf, 2018b).

#### 4. Relate Design Attributes to Girls' Engagement (Differs in reanalysis)

The goal of the original EDGE research was to winnow the large set of promising female-responsive design attributes down to the best ones for engaging girls in a STEM learning environment in order to influence practice. To that end, the analysis required a high bar of confidence, surfacing only those DAs that consistently, positively, and significantly related to girls' engagement at STEM exhibits. The best DAs from the girls' data were then checked against the boys' data for any ill effects on boys' engagement; none were found. The high bar allowed EDGE to advocate for changing exhibit designs to be more inclusive for girls. The present reanalysis had the goal of describing the status

quo, not making recommendations for practice, and thus employed a different process for analyzing the data (explained below).

### **Reanalyzing the Data: State of the Field and Children's Engagement**

The goal of the present reanalysis is to provide a snapshot of the field's current state for researchers— how do the DAs relate to various measures of children's engagement, and how common or rare are they? This information will help researchers begin to distinguish designs that offer low returns on investment from those that presently show promise for visitor learning, in order to conduct studies that push the field beyond the status quo. Additionally, our survey of DAs at exhibits in the content areas of physics, technology, engineering, math, time, and perception begins to document the prevalence of science museum exhibit designs in those domains. To our knowledge, no synthesis study about exhibit design has been conducted to date with a large number of design attributes across many exhibits and multiple institutions. Still, we must keep in mind that the exhibits were selected based on convenience, and overrepresented at the larger institutions (particularly the West Coast museum). Further, DAs were developed and selected based on a review of girls' STEM learning, rather than broader notions of exhibit designs. For example, none of the DAs were concerned with wheelchair accessibility, live specimens, or audio labels. Our reanalysis of the EDGE data applies to only a subset of all possible exhibit design attributes.

### **Changes From EDGE**

Because the research questions differed substantially from the original project, we changed various aspects of both the dataset and the statistical analyses employed in

EDGE. These changes affected our level of confidence in the outcomes and created some differences between EDGE and the reanalysis, making direct comparisons difficult and unwise. Table 3 shows the differences between the original analyses and the current reanalysis. Perhaps the most important change was that the original analysis employed inferential statistics (with  $p$  values) in order to make claims about which design attributes would best support girls' engagement. That project was looking for consistently strong positive relationships with high confidence in order to affect practice. Based on the power analysis for that original study, the inferential statistics required a minimum sample of 50 exhibits representing each DA to detect small effects, which are common in museum studies (Cohen, 1992; Wilson, Voorhis, & Morgan, 2007).

Table 3

*Differences between Original Analysis and Reanalysis*

	EDGE	Reanalysis of dataset
Research question	What are the best design attributes for engaging girls at exhibits?	What is the status quo regarding use and design of exhibits?
Focal group	Girls	Children: girls and boys combined
Analyses	Correlation followed by focused regression with semi-partial correlations. Includes inferential statistics and $p$ values.  Rationale: Identifies the best design attributes with high confidence and ensures that no results are driven by a single institution.	Semi-partial correlation. Solely descriptive statistics without $p$ values.  Rationale: Provides current snapshot of design attributes at STEM exhibits. Identifies promising areas for future R&D.
Minimum number of exhibits with an attribute needed	50  Rationale: Inferential regression analyses recommend a large sample size.	3  Rationale: Museum studies typically require a minimum of three exhibits for generalization.
Decisions regarding design attributes	<ul style="list-style-type: none"> <li>• Included only girl-friendly DAs.</li> <li>• DAs rated on a scale were categorized into presence or absence based on the data.</li> <li>• Choices among combinations and variations of DAs were made based on the strength of their relationships with the four engagement measures.</li> </ul> Rationale: The aim was to identify the strongest and best DAs for engaging girls.	<ul style="list-style-type: none"> <li>• Included non-girl-friendly and girl-unfriendly DAs (e.g., competition).</li> <li>• DAs rated on a scale were categorized into presence or absence based on an across-the-board decision of medium/low vs. high.</li> <li>• Choices among combinations and variations of DAs were made prior to analysis based on across-the-board decisions. Many redundant DAs were retained to show how the relationship to engagement measures differs depending on subtle variations in their operationalization.</li> </ul> Rationale: The intent was to provide a full picture; such comprehensiveness requires consistency and ease of interpretation.
Controls for institution and crowdedness	Yes	Yes
Describes negative relationships	No Rationale: Analysis identifies best attributes, so negative relationships are dropped.	Yes Rationale: Snapshot requires full picture (positive and negative correlations).
Inter-rater reliability of highest engagement level	Kappa Rationale: We wanted to have greater confidence in the claims so we set a higher bar for reliability by controlling for chance agreement. This affected the number of exhibits in the analysis.	Percentage agreement Rationale: This allowed us to include the majority of the exhibits in the analysis. Because we are not making strong claims, we can learn from the less rigorous approach.

In contrast, the current reanalysis attempts to document the landscape without generalizing beyond the sample, acknowledging that such descriptive results may be due to chance. This approach allows us to include rare DAs (present at a minimum of only three exhibits) and DAs that correlate with poor outcomes. Unlike the original analysis, we did not conduct significance tests, because doing so on over 60 DAs across each of the four engagement measures would have increased the likelihood of a Type I error to perilously high levels. Instead, we calculated correlations between each DA and engagement measure, which differed in precision depending on the number of exhibits with that design attribute. This allowed us to identify areas of promising research, rather than make inferences about DAs to embrace or avoid.

## **Results of Reanalysis**

### **Baseline levels of Engagement**

How did the exhibits tend to engage children, regardless of specific design attributes? Before delving into the DAs and their relationships with engagement, we present descriptive statistics for the four engagement measures across all exhibits and institutions in the study. These statistics serve two functions. First, they provide baselines for the remaining results that describe associations between DAs and engagement measures. Second, the descriptive statistics offer a portrait of typical engagement levels at physics, engineering, math, time, and perception exhibits among 8–13 year olds at three representative museums. However, it is important to note that these results are more representative of engagement at the institutions with more exhibits (i.e., West Coast followed by Midwest). Indeed, over two-thirds of the exhibits were located at the West Coast museum.

Table 4 shows the engagement baselines for children averaged across all exhibits. The mean percentage of children who used an exhibit was 22%; the mean time children spent at an exhibit was 62.5 seconds; on average, the percentage of children who returned to an exhibit they had already used was 15%; and the average percentage of children who reached the highest level of engagement at an exhibit was 88%.<sup>5</sup>

---

<sup>5</sup> The large percentage of children reaching the highest level of engagement represents restricted variability (range restriction between 70% and 100%), and may therefore underestimate some of the correlations between design attributes and High Engagement Level (presented in the next section).

Table 4

*Descriptive Statistics of Average Exhibit Engagement for Children*

Engagement measure	Mean (SD)	95% CI	Minimum	Maximum
Use at exhibits (%)				
Overall	22% (16%)	20% - 24%	2%	98%
Southwest	70% (16%)	.62% - .78%	45%	98%
Midwest	24% (12%)	.21% - .27%	7%	63%
West Coast	18% (8%)	.16% - .19%	2%	44%
Time Spent (seconds)				
Overall	62.5 (40.60)	57.91 - 67.11	9.2	273.9
Southwest	76.99 (39.42)	57.39 - 96.59	25.99	162.94
Midwest	83.75 (52.22)	71.3 - 96.2	9.24	270.76
West Coast	54.30 (32.89)	49.86 - 58.75	12.48	273.94
Return Visits (%)				
Overall	15% (9%)	14% - 16%	0%	60%
Southwest	29% (15%)	.22% - .36%	10%	63%
Midwest	14% (13%)	.13% - .18%	0%	50%
West Coast	13% (9%)	.12% - .14%	0%	50%
High Engagement Level (%)				
Overall	88% (10%)	86% - 89%	36%	100%
Southwest	82% (7%)	.79% - .86%	70%	94%
Midwest	83% (12%)	.80% - .86%	36%	100%
West Coast	89% (8%)	.88% - .91%	46%	100%

*Note.* Descriptive statistics of participant behavior involved averaging across all exhibits. n=301 exhibits overall; n=18 exhibits at the Southwest museum; n=70 exhibits at the Midwest museum; and n=213 exhibits at the West Coast museum.

There was variability across institutions, perhaps most strikingly at the Southwest museum, which had the smallest space by far. With only 18 exhibits, that museum saw a much larger percentage visitors using and returning to each exhibit on average. Visitors also spent less time on average at exhibits in the large, West Coast museum. The percentage of children reaching a High Engagement Level was similar across museums.

### **Design Attributes and Children's Engagement**

The current analysis calculated semi-partial correlations between DAs and children's engagement for each of the four engagement measures. Semi-partial correlations show the strength of the relationship between each design attribute and an engagement measure (i.e., the size of the effect), over and above institutional differences and the effects of exhibit crowdedness on the engagement measures. Partialing out institutional differences helps us account for the differences in results that can be seen in Table 4 (potentially due to museum size, pedagogical approach, visitor demographics, location, and number of exhibits in the space). We provide the full list of attributes and their semi-partial correlations with each type of engagement in Table 5.

To illustrate how we might use these effect sizes to inform future research, we present a specific cutoff in the table: greater or less than 0.1. Our cutoff is based on the standard rule of thumb used in social science research, where 0.1 is considered a small but meaningful effect, 0.3 is considered moderate, and 0.5 is considered to be a large effect (Cohen, 1988, 1992). Further, a group of advising practitioners expressed an interest in the relationships that showed this size of effect, when we discussed the relationship in terms of the baseline levels. For example, a correlation of 0.1 translates to approximately a 4-9% increase in exhibit use (c.f., the baseline average of 22%); approximately a 11-13 second increase in time spent (c.f., the baseline average of about one minute); approximately a 2-5% increase in return visits (c.f., the baseline average of 15%); and approximately a 3-4% increase in highest engagement level reached (c.f., the baseline average of 88%).<sup>6</sup>

---

<sup>6</sup> These ranges are based on the regression slopes for effect sizes of .10 to .14.

For the purpose of illustration, we identify good, neutral, and poor semi-partial correlations using the 0.1 cutoff in the table. A “good” semi-partial correlation means that the presence of the design attribute was associated with higher engagement. In technical terms, the semi-partial correlation was 0.1 or greater. A “poor” semi-partial correlation means a negative correlation: The presence of that design attribute was related to lower engagement (the semi-partial correlation was -0.1 or less). A “neutral” semi-partial correlation means the design attribute was not correlated with the engagement measure (semi-partial correlation was between 0.1 and -0.1). We report the detailed semi-partial correlations to allow readers to interpret the size of the effect, choose their own cutoffs if they wish, and employ the results in power analyses to determine sample sizes necessary for future research. To sort our table, we calculated the ranking for each design attribute within each engagement measure, averaged across the four rankings, sorted by the resulting mean rank (highest to lowest), and finally adjusted for legibility.

The table also shows the prevalence of each DA by indicating the number and percentage of exhibits incorporating that attribute. The prevalence of each DA across exhibits helps us understand not only whether such designs arise frequently or rarely across the three museums, but also how reliable or precise is our measure of their relationship to engagement. That is, the fewer exhibits with that attribute, the less precision in estimating the relationship (or the more heavily the finding relies on those specific exhibits rather than the design attribute generally). Finally, in order to put these results in the context of the larger EDGE findings, the exhibit design attributes that were

the most important for engaging girls are marked with this symbol: .<sup>7</sup>

---

<sup>7</sup> Although the table highlights the nine DAs that were discovered as helpful to girls in EDGE, several of the DA variables differ between the two analyses, and these differences affect the sample sizes. For example, EDGE explored combinations of variables, and one of the top DAs for designing for girls was the combination variable: *exhibit has multiple sides or stations*. The reanalysis reported here maintained these two DA variables as separate. Another example includes *exhibits' look-and-feel is homey, personal, homemade* which was collected using a three-point scale; EDGE split the variable into high/medium vs. low, while the reanalysis split the variable into high vs. low/medium (based on the best approach for answering the respective research questions).

Table 5  
*Relationships between Design Attributes and Children's Engagement*

Design Attribute	Use	Time Spent	Return Visits	High Eng. Level	Exhibits with DA (max = 301)	% of exhibits with DA
 Has multiple stations	.09	.23	.15	.1	44	15%
Phenomenon can be experienced by 2 or more people at the same time	.09	.13	.1	.12	240	80%
Label suggests three or more distinct activities	.02	.18	.14	.12	25	8%
 Space to accommodate three or more people	.17	.16	.24	.07	170	56%
Allows for more than one set of hands or bodies	.12	.14	.15	.01	224	74%
Designed for multiple players to use without interfering with each other	.2	.17	.23	-.1	94	31%
Exhibit has a competitive feel	.07	.28	.11	.07	17	6%
 Exhibit is open-ended	.05	.31	.16	.04	132	44%
One or more visitors can lose	.04	.15	.17	.03	4	1%
 Can be used from multiple sides	.24	.07	.26	.08	94	31%
Label invites visitors to work together	.12	.02	.11	.07	20	7%
 Visitors can watch others from afar to preview what to do	.1	-.01	.14	-.03	232	77%
 Exhibit includes at least one familiar object	.05	.13	.12	-.001	97	32%
Title suggests what to do at the exhibit	-.11	.22	-.01	.17	147	49%
Interactive elements are observable prior to use	-.01	.12	-.03	.12	225	75%
 Exhibit includes a use drawing	.03	.08	.07	.13	98	33%
 Exhibit includes at least one image of a person	.05	.08	.09	.18	126	42%
Title suggests what the exhibit does	.09	.01	0	.14	223	74%
Exhibit has a label	.06	.01	-.02	.21	291	97%
Title is legible from afar	.03	.02	.01	.12	256	85%
Label has bright, prominent color	0	.09	-.06	.14	70	23%
Label asks at least one open-ended question	.02	0	-.01	.13	148	49%
Includes two or more required roles	.11	.04	.09	.05	14	5%
 The exhibits' look-and-feel is playful, whimsical, or humorous	.06	.06	.15	.001	83	28%
Exhibit includes any text about an ambiguously gendered (OR non-gendered) STEM professional	.03	.06	.1	-.002	21	7%
Label provides history of the exhibit or the phenomenon	.09	-.04	.11	-.03	60	20%
Seating for two or more	-.03	.32	.04	-.01	56	19%

 = EDGE DA

Good
Neutral
Poor

Design Attribute	Use	Time Spent	Return Visits	High Eng. Level	Exhibits with DA (max = 301)	% of exhibits with DA
Exhibit involves self-expression or authorship	-.08	.36	.08	-.05	38	13%
Prior visitors' work is visible from afar	-.07	.37	.07	-.08	26	9%
Exhibit has seating	-.08	.16	-.07	.06	147	49%
Exhibit includes any image of a female	.06	0	.09	.07	38	13%
Exhibit includes any image of a male	.03	.04	.06	.08	52	17%
The overall color palette for the exhibit is: Public signage or safety colors (red/orange/yellow with gray/black)	.02	.02	.05	.09	43	14%
The exhibits' look-and-feel is industrial, slick, bulky, or indestructible	.05	-.04	.05	.09	121	40%
The exhibit is designed to reflect a visitors' self-image	.05	.02	.05	.02	34	11%
The overall color palette for the exhibit is: Neutral colors (white, black, gray, brown, wood, metal, mirror, clear)	-.01	.01	.08	.07	138	46%
Label images are informal, hand-drawn, and casual, rather than precise or formal	.03	-.06	.05	.07	33	11%
Exhibit has bright, prominent color in the secondary elements (such as cabinetry or hardware)	.07	-.02	.08	-.03	17	6%
Label encourages visitors to compare with others	0	.02	.02	.06	10	3%
Exhibit includes any image of a STEM professional	.03	-.04	.01	.05	6	2%
The exhibits' inner workings are visible	.03	.01	-.03	.05	100	33%
Exhibit includes any text about a female STEM professional	-.03	.02	.08	0	4	1%
The overall color palette for the exhibit is: Subdued, desaturated colors (pink, burgundy, rust, pale green, olive, aqua)	.02	.04	-.05	.02	23	8%
Exhibit includes any text about a male STEM professional	.03	-.08	.04	.01	26	9%
Label text tone is informal and friendly, rather than formal or academic	-.02	.09	-.002	-.04	33	11%
Visitors can watch others at the exhibit to preview what to do	-.09	.06	.03	-.01	287	95%
Exhibit provides opportunity to use small motor skills	-.07	.08	-.07	.04	264	88%
Label describes how the exhibit phenomenon is related to social issues for humans, animals, or the environment	-.01	-.09	-.01	.01	42	14%
 The exhibits' look-and-feel is homey, personal, homemade, or delicate	-.02	-.03	-.05	-.02	25	8%

 = EDGE DA

Good
Neutral
Poor

Design Attribute	Use	Time Spent	Return Visits	High Eng. Level	Exhibits with DA (max = 301)	% of exhibits with DA
Exhibit uses a familiar object in an unfamiliar way	.03	- .12	.01	-.02	24	8%
Label describes how the exhibit phenomenon is used in the real world	.004	- .13	-.02	-.02	157	52%
Exhibit has bright, prominent color in the phenomenon or interactive elements	-.03	.09	-.04	-.1	113	38%
Exhibit has embedded story or narrative	-.06	-.01	-.05	-.13	7	2%
The overall color palette for the exhibit is: Vivid, saturated colors (red, orange, yellow, green, blue, purple)	-.02	-.05	-.1	-.15	97	32%



EDGE DA

Good

Neutral

Poor

To further exemplify how we might use these data to guide our own future research, we present four additional tables (6-9) that aid interpretation. These tables, one for each of the engagement measures, present only the DAs that had a notable positive semi-partial correlation ( $sr \geq 0.1$ ) with that engagement measure. The tables also highlight the prevalence of each DA across exhibits, helping us understand, again, how reliable or precise is our measure of their relationship to engagement. We considered DAs to be “widespread” if they were found at more than 100 exhibits (>33%) in our sample; “fairly common” DAs were present at 30-100 exhibits (10-33%); and *rare* DAs were found at less than 30 exhibits (<10%). Finally, for ease of reading and interpretation, we classified disparate DAs into three groups based on three aspects of exhibits: exhibit labels, exhibit look-and-feel, and exhibit interactions. Table 5 provides the full data for readers to create their own groupings and cutoff criteria to support their own explorations. For each exhibit aspect, some design attributes were notably correlated, and some were not. The following four tables only present those that were positively correlated ( $sr \geq 0.1$ ). For ease of interpretation, each table has been sorted by

these groupings (in the order listed above) and prevalence (from rare to fairly common to widespread).

Table 6

*Design Attributes Positively Correlated with Exhibit Use*

Design Attribute	Semi-partial correlation	Exhibit n	Prevalence
Labels			
Label invites visitors to work together	0.12	20	Rare
Interactions			
Includes two or more required roles	0.11	14	Rare
👩 Can be used from multiple sides	0.24	94	Fairly Common
Designed for multiple players to use without interfering with each other	0.2	94	Fairly Common
👩 Space to accommodate three or more people	0.17	170	Widespread
Allows for more than one set of hands or bodies	0.12	224	Widespread
👩 Visitors can watch others from afar to preview what to do	0.1	232	Widespread

*Note.* The girl symbol 👩 indicates DAs that were also found to be successful in engaging girls in the EDGE study.

Table 7

*Design Attributes Positively Correlated with Time Spent*

Design Attribute	Semi-partial correlation	Exhibit n	Prevalence
Labels			
Label suggests three or more distinct activities	0.18	25	Rare
Title suggests what to do at the exhibit	0.22	147	Widespread
Look & Feel			
👧 Exhibit includes at least one familiar object	0.13	97	Fairly Common
Interactions			
Prior visitors' work is visible from afar	0.37	26	Rare
Exhibit has a competitive feel	0.28	17	Rare
One or more visitors can lose	0.15	4	Rare
Exhibit involves self-expression or authorship	0.36	38	Fairly Common
Seating for two or more	0.32	56	Fairly Common
👧 Has multiple stations	0.23	44	Fairly Common
Designed for multiple players to use without interfering with each other	0.17	94	Fairly Common
👧 Exhibit is open-ended	0.31	132	Widespread
Exhibit has seating	0.16	147	Widespread
👧 Space to accommodate three or more people	0.16	170	Widespread
Allows for more than one set of hands or bodies	0.14	224	Widespread
Phenomenon can be experienced by 2 or more people at the same time	0.13	240	Widespread
Interactive elements are observable prior to use	0.12	225	Widespread

*Note.* The girl symbol 👧 indicates DAs that were also found to be successful in engaging girls in the EDGE study.

Table 8

*Design Attributes Positively Correlated with Return Visits*

Design Attribute	Semi-partial correlation	Exhibit n	Prevalence
Labels			
Label suggests three or more distinct activities	0.14	25	Rare
Label invites visitors to work together	0.11	20	Rare
Exhibit includes any text about an ambiguously gendered (OR non-gendered) STEM professional	0.10	21	Rare
Label provides history of the exhibit or the phenomenon	0.11	60	Fairly Common
Look & Feel			
👧 The exhibits' look-and-feel is playful, whimsical, or humorous	0.15	83	Fairly Common
👧 Exhibit includes at least one familiar object	0.12	97	Fairly Common
Interactions			
One or more visitors can lose	0.17	4	Rare
Exhibit has a competitive feel	0.11	17	Rare
👧 Can be used from multiple sides	0.26	94	Fairly Common
Designed for multiple players to use without interfering with each other	0.23	94	Fairly Common
👧 Has multiple stations	0.15	44	Fairly Common
👧 Space to accommodate three or more people	0.24	170	Widespread
👧 Exhibit is open-ended	0.16	132	Widespread
Allows for more than one set of hands or bodies	0.15	224	Widespread
👧 Visitors can watch others from afar to preview what to do	0.14	232	Widespread
Phenomenon can be experienced by 2 or more people at the same time	0.10	240	Widespread

*Note.* The girl symbol 👧 indicates DAs that were also found to be successful in engaging girls in the EDGE study.

Table 9

*Design Attributes Positively Correlated with High Engagement Level*

Design Attribute	Semi-partial correlation	Exhibit n	Prevalence
Labels			
Label has bright, prominent color	0.14	70	Fairly Common
👧 Exhibit includes a use drawing	0.13	98	Fairly Common
Exhibit has a label	0.21	291	Widespread
👧 Exhibit includes at least one image of a person	0.18	126	Widespread
Title suggests what to do at the exhibit	0.17	147	Widespread
Title suggests what the exhibit does	0.14	223	Widespread
Label asks at least one open-ended question	0.13	148	Widespread
Title is legible from afar	0.12	256	Widespread
Interactions			
👧 Has multiple stations	0.10	44	Fairly Common
Phenomenon can be experienced by 2 or more people at the same time	0.12	240	Widespread
Interactive elements are observable prior to use	0.12	225	Widespread

*Note.* The girl symbol 👧 indicates DAs that were also found to be successful in engaging girls in the EDGE study.

## Discussion

This reanalysis highlights the relationships between existing science museum exhibit designs and children's engagement. Importantly, current designs represent the status quo, a state of affairs in museums that unfortunately promote inequities, often advantaging white, middle-class, and male ways of learning (Dawson, 2014a, 2014b; Feinstein, 2017). Rather than reify present circumstances by using the results to recommend some existing designs over others, we present the findings to identify productive subjects of future research.

The reanalysis of the EDGE data provides two kinds of results. The first offers baseline data. By looking across all of the exhibits in the study, we can provide an

average or typical engagement for all four measures. By comparing those results to prior studies that used similar measures, we can add to the understanding the field holds about how children engage with science museum exhibits.

The second kind of result presents children's engagement data for a variety of design attributes along with how common or rare those attributes were. The reanalysis of the EDGE data indicates that several design attributes were positively correlated with children's engagement, and a smaller number correlated with negative engagement for children. Again, rather than prescribe use of the former and avoidance of the latter, we suggest that the results guide future research and evaluation regarding exhibit design.

### **Baseline Statistics**

Providing the baseline levels for each area of engagement can help researchers set expectations and clarify impacts. Below we share what we learned from the reanalysis about children's general engagement with STEM exhibits. We also point out instances when our data align with prior research (i.e., replicate prior results).

Using an exhibit is a pre-requisite to any kind of engagement or deeper learning opportunities. Our analysis found that, averaging across exhibits, only 22% of children used a typical exhibit. This lies at the low end of prior research in science exhibit "attracting power," the ability of an exhibit to entice museum-goers to stop and try it. For example, Yalowitz and Bronnenkant (2009), describing a study of 44 exhibits in an aquarium exhibition, reported that the average exhibit attracted 41% of general visitors; Sandifer (2003) found that 38% of visitors of all ages stopped at the average science museum exhibit among 61 exhibits spread across two exhibitions; finally, Bosivert & Slez (1995) tracked visitors of all ages as they moved among 80 exhibits and discovered

that only 19% stopped at the average exhibit. Perhaps there is an inverse relationship between number of exhibits and attracting power, or perhaps other differences in context, exhibits, or visitors' ages bear responsibility for the variation. However, even if science museum exhibits on average are used by only 22% of children, this could represent a large number of people, especially if we extend that finding to approximate usage by visitors of all ages. At the Exploratorium whose annual attendance is 800,000, 22% of visitors would translate into the average exhibit being utilized by 176,000 visitors per year. This suggests that over the life of a science museum exhibit, millions of people experience its marvels. No wonder Frank Oppenheimer, founder of the Exploratorium, suggested that every exhibit be iteratively improved over time to maximize its effectiveness (Oppenheimer, 1986).

Perhaps one of the most-studied aspects of the exhibit encounter has been holding time (Sanford, 2010; Serrell, 1998, 2010; Yalowitz & Bronnenkant, 2009). Across multiple studies conducted at small numbers of science museum exhibits, the average visitor or group was found to spend about one minute interacting with the average exhibit (Falk, 1983; Humphrey & Gutwill, 2005; Sandifer, 2003). The present analysis found the same result with children, providing robust replication and setting this baseline firmly in place. As new projects develop or study innovative exhibits to deepen engagement, especially for underserved audiences, we suggest they continue to use this result as a benchmark beyond which to push their own development.

We found that 15% of children returned to an exhibit after using it. We believe return visits are important because they reflect a visitor's deep enough interest to experience something again, use something after observing someone else use it, or bring

someone else over to show them an exhibit. Again, assuming that children and adults behave similarly in this regard, this would mean that approximately 120,000 people annually use the average exhibit more than once during a visit to the Exploratorium. Although such findings require replication with adults, they provide a benchmark for exhibits seeking to offer rich sets of activities so millions of returning users can deepen their learning experiences over the lifetime of the exhibit.

Lastly, we found that 88% of children reached the highest level of engagement in our outcome measure. That is, the vast majority of children were actively repeating or varying inputs or sharing their experience with others while at the exhibit. This finding fluctuated little across the three museums. This may suggest that most interactive, hands-on science museum exhibits successfully promote basic inquiry behaviors. Still, further research on rates of minimum inquiry may be warranted in other types of informal learning institutions, such as history, natural history, art, and cultural museums.

Notably, participants behaved differently at the three museums (as shown above in Table 4). The results seem to underscore the importance of environmental design. For example, the smaller Southwest institution had the highest mean, minimum, and maximum percentage of Use and Returns; the Midwest institution had the midranges for Use and Returns; and the West Coast museum's Use and Return results were lowest. Scrutinizing the environmental design (Figure 1), we see that when fewer exhibits were presented, children spread their attention across the exhibits available to them, rather than using a fixed percentage of exhibits in a given area. (The average length of time of a museum visit tends to remain steady for large and medium size museums (R. Auster, personal communication, February 16, 2018).) Regarding Time Spent, the West Coast

museum had the lowest average across their exhibits. This may be partially due to a larger number of exhibits in the study space or a higher density of exhibits. Finally, when we look to the High Engagement Level results, we find that once again the smaller institution had a much higher minimum average percentage of children who repeat, vary, or talk about their experience. Again, this lack of variability at the medium-sized museum may be due to visitors trying get the most out of a small number of exhibits, or the environmental design of a small room allowing for deeper focus.

The differences across institutions in the four engagement results may be related to prior research investigating the effects of walls on visitors' exhibit experience. In two studies conducted at different institutions, the presence of walls, which created a smaller space with the same exhibit density, increased holding time (Adamson, 2008; Allen, 2004). (However, one study found that walls also increased attraction power while the other study found the opposite effect.) In our view, the impact of environmental cues on exhibit experience remains a rich domain for future research.

### **Informing Research to Advance the Field**

This study provides ample information regarding the relationship between each DA and each engagement measure, along with the prevalence of each attribute in the data set. Next we detail several examples of how we might use these data to guide our own research agendas. We encourage others to explore similar paths and use the data to generate their own questions to further the field. Several of the findings support prior research conducted in this arena, while other aspects of the results seem to raise new questions.

**Well-studied design attributes/replication.** Several of the DAs that have been found effective in prior research received further support here. Specifically, multiuser / multistation (i.e., *exhibit has multiple sides or stations*) designs showed positive correlations across measures. Multiple option (i.e., *label suggests three or more activities*) and multi-outcome (i.e., *exhibit is open-ended*) attributes also proved themselves valuable in our study. As these design attributes were found effectual in a variety of studies (Borun et al., 1998; Dancstep et al., 2015; Humphrey & Gutwill, 2005; Perry, 2012; Sandifer, 2003; Serrell, 1998, 2010), they may be viewed as sound design ideas for typical science museum audiences without further research.

In addition, the current study suggests that *including at least one familiar object*, employing *labels that clearly suggest what to do*, and *allowing visitors to watch others from afar to preview what to do* were all positively related to engagement. All three of these relate to Perry's design principles as well as concepts from label researchers like Beverly Serrell (Perry, 2012; Serrell, 2015). Familiar objects and visitors watching others from afar to preview what to do also emerged from the EDGE analyses as important to girls. Discovering that they were positive for boys as well underscores the perspective that sometimes designing for accessibility with one underserved group in mind can sometimes improve experience for many (Reich, 2005). Based on past and present results, we believe that familiar objects, clear labels, and orientation from afar probably need little further research with existing science museums audiences. Of course, results from this reanalysis represent the status quo; further research is needed with underserved and intersectional audiences.

**Common but understudied.** Two DAs arose that are common, but have not been studied much (to our knowledge): *label has bright, prominent color* and *exhibit includes at least one image of a person*. In our experience, color sometimes becomes an issue of contention among designers; while some see real power in the inclusion of color, others voice concern that color distracts from the phenomenon (unless used strategically to highlight it). We have heard from many designers that they want more solid data to help inform these decisions. The second design attribute, using an image of a person, emerged in EDGE as an important design attribute for engaging girls and also has been found elsewhere to contribute to a feeling of welcome for girls (Crowley, Callanan, Tenenbaum, & Allen, 2001). Given the result that both label attributes correlated with engagement for boys and girls suggests that more research is warranted.

**Rare but promising design attributes.** For three promising DAs that appeared in exhibits only rarely, it is unclear whether the positive results would hold with a larger sample size: *exhibit has a competitive feel* (N=17 exhibits); *one or more visitors can lose* (N=4); and *label suggests three or more distinct activities* (N=25). The first two seem related. In a prior study, competition was associated with longer holding times and more return use for children (Garibay, 2013). Perhaps the competitive activities were exciting and demanding, prompting children to keep playing until they could win. However, this result challenges prior research on girls' preferred ways of learning and knowing (National Science Foundation, 2003; Peterson & Fennema, 1985; Rosser, 1991; Taylor, 2005) and thus seems an important area for further study and deeper understanding (more on this below). *Label suggests three or more distinct activities* fits well with multi-option, multi-outcome designs. However, we seldom found exhibits with this label attribute,

indicating that further study could inform the field as to whether this simple label change should be utilized more frequently.

**Negative associations.** A small number of DAs actually showed negative correlations with measures of engagement. The EDGE study eliminated any DAs with negative correlations, because the goal was to discover the best designs for engaging girls (those that were consistently, strongly, and positively related). However, when we consider how this full data set can inform our future research agenda, we believe that common, negative design attributes require further investigation. Three DAs fell into this category: *Exhibit has bright, prominent color in the phenomenon or interactive elements; overall color palette for the exhibit is vivid, saturated colors (red, orange, yellow, green, blue, purple); and label describes how the exhibit phenomenon is used in the real world.* The first two seem related, but contradict an earlier finding. Recall that when the *label contained bright, prominent colors*, children tended to show higher engagement at the exhibits. The opposite occurred when the *main aspects of the exhibit itself employed bright colors*. This inconsistency begs for more research on the use of bright color on exhibits and labels. The last commonly found DA that negatively correlated with engagement also surprised us. Making real world connections in labels is a widespread practice, supported by research and practical wisdom (e.g., Ansbacher, 1999; Atkins, Velez, Goudy, & Dunbar, 2009; Perry, 2012; Serrell, 2015). Why did it correlate with less holding time for children? Does a major difference exist between children and adults in this respect, or might the finding be related to label length (Bitgood & Gregg, 1986; Bitgood & Thompson, 1989)? We hope for additional study.

**Taking a larger view.** Stepping back from the individual DAs, we see a larger pattern in the results, particularly in the positive correlations between DAs and engagement. Exhibit Use, Time Spent, and Return Visits seemed to correlate positively with a remarkably small, stable set of design attributes. Those designs tended to match what were found in the PISEC, APE, and other studies: Science exhibits promote engagement well when multiple stations offer groups numerous, non-interfering options that can lead to multiple outcomes (Borun et al., 1998; Humphrey & Gutwill, 2005; Sandifer, 2003). This result further supports the previously developed idea that science museum visitors prefer to stay with their friends and family while using exhibits and will explore more deeply when exhibits provide greater complexity (Falk & Dierking, 2000; Humphrey & Gutwill, 2005; Perry, 2012; Sandifer, 2003).

In contrast to the other three engagement measures, a high engagement level score (repeating, varying, or talking about an experience while at an exhibit) seemed to correlate with a different set of design attributes. Label designs, more than exhibit designs, appeared to be important for promoting manipulation of variables or communication among visitors at exhibits. *Titles that describe what to do at an exhibit or what the exhibit does* and the *labels that include use drawings* were related to a high engagement level. These kinds of label designs have been tangentially studied before (Humphrey & Gutwill, 2005), but the field would benefit from more rigorous research.

**Differences from EDGE.** These data were originally collected to address issues of disparity in science centers that echo broader equity and inclusion issues for women and girls in STEM (Borun, 1999; Diamond, 1994; Girls Inc., 2004; Greenfield, 1995; Hill, Corbett, & St. Rose, 2010; Kremer & Mullins, 1992; National Science Foundation,

2003, 2007; Verheyden, 2003). As described earlier, the results from the EDGE study warrant a higher level of confidence and allow us to advocate for design attributes that best support girls' engagement. The results from the reanalysis differ from the EDGE in many ways. These differences are due partly to the inclusion of boys' data, and could, therefore, serve to reify existing disparities. Differences also arise from the changes in the analytic approach—exploring the status quo in the reanalysis allowed us to highlight negative relationships and investigate designs represented by fewer exhibits, but limits any claims of generalizability beyond the sample. We discuss a few of these discrepant outcomes below.

Some of the design attributes that were strongly related to engagement in the reanalysis differ from what we found in EDGE, or even run counter to existing research regarding females' STEM learning preferences. For example, regarding the DA of *label has bright, prominent color*, EDGE found that the relationship with girls' behavior was not significant and therefore may be due to chance. However, the reanalysis did not employ inferential statistics, so we do not know if the positive relationship found herein is significant or simply due to chance. Another example includes the competitive attributes—*exhibit has a competitive feel* and *one or more visitors can lose*—which were positively related to children's engagement in the reanalysis. EDGE did not analyze the impact of either of these competition attributes because too few exhibits represented them. As mentioned earlier, competition as a design approach may negatively impact females (National Science Foundation, 2003; Peterson & Fennema, 1985; Rosser, 1991; Sammet & Kekelis, 2016; Taylor, 2005; Wellcome Trust, 2013). However, the impact may depend on the type of competition: One of the project's expert interviewees

suggested that girls may enjoy group competition, but not wish to be pitted against one another in individual competition (Koke, 2007). Moreover, issues with competition may no longer hold in informal learning environments: The EDGE project's Girl Advisory Committee expressed enjoyment at an exhibit that incorporated competition. These two attributes represent complicated relationships between design and visitor engagement. We recommend studies that try to tease out the effects of variations in these designs, in addition to keeping a keen eye out for negative outcomes for underrepresented audience members, such as females.

Finally, four of the design attributes that were most successful for engaging girls in EDGE did not consistently rise to the top in the reanalysis: (1) *exhibits' look-and-feel is playful, whimsical, or humorous*; (2) *exhibits' look-and-feel is homey, personal, homemade, or delicate*; (3) *exhibit includes a use drawing*; and (4) *exhibit includes at least one image of a person*. We advocate for incorporating the EDGE DAs in order to push museum practice and create learning environments that are welcoming to a broader group of visitors, rather than remaining rooted in the status quo.

### **Conclusion**

This reanalysis of EDGE data has provided a snapshot of the status quo of science museum exhibit designs. Although the dataset was quite large compared to those of previous studies, the exhibits in the current research came from only three representative science museums (using a convenience sample that overrepresents the larger institutions), focused on STEM content that excluded biology and social science, and were originally investigated with coding schemes for design attributes related to females' ways of learning and knowing (Dancstep & Sindorf, 2018a). Further, while the team increased

data collection efforts on free days and Latino family program days, the sample included relatively few participants from non-dominant economic or cultural communities. The sample also lacked participants with non-binary genders, perhaps due to the comparatively young age of children in the study. These results may not extend to adults or even children in other age ranges. While we value ecologically valid studies, research within museums is often limited to certain audiences. We recommend that future research explore mixed and alternate methodologies that amplify non-dominant voices. More work is needed to understand exhibit design and the intersection of multiple identities including gender, economic, ethnic, and STEM.

Given these limitations, along with the reliance on descriptive rather than inferential statistics, we recommend caution in applying the results of this reanalysis to practice. In addition, the data presented here represent the current state of the art in exhibit design, which may actually inhibit engagement for disenfranchised learners (Dawson, 2014a, 2014b; Feinstein, 2017). Simply implementing the positive design attributes and eradicating the negative ones could reify existing inequities. Instead, we present these data for researchers as one resource for determining where the field's investigative energy is best spent. To that end, we have shown examples where support already exists, and where we see promising areas for further research. Other researchers are encouraged to explore the data to inform their own next steps. The present reanalysis has shown us, and hopefully others in the museum field, where to begin.

### **Acknowledgements**

We appreciate the thoughtful contributions by the project team, including researchers, exhibit and content developers, graphics and editorial staff, and project management.

This research would not be possible without the museums and visitors that agreed to participate. We thank the members of our esteemed Advisory board: Todd Bodner, Judy Brown, Lynn Dierking, Cecilia Garibay, Kris Gutierrez, Judy Lee Haworth, Laura Huerta Migus, and Don Norman. We developed this article with generous feedback from reviewers, several EDGE Workshop participants, and museology graduate students at the University of Washington (special thanks to Dr. Jessica Luke and Dylan High). This material is based upon work supported by the National Science Foundation under Grant Number DRL-1323806. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

- Adamson, B. (2008). *The relationship between exhibition environments and visitors' behaviors*. Paper presented at the Visitor Studies Association, Houston, TX.
- Allen, S. (2004). *To partition or not to partition: A study of the impact of walls on visitor behavior at an exhibit cluster*. Paper presented at the Visitor Studies Association, Albuquerque, NM.
- Ansbacher, T. (1999). Experience, inquiry and making meaning. *Exhibitionist*, 18(2), 22-26.
- Association of Science-Technology Centers. (2016). ASTC Statistics Package. Washington, D.C.
- 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.
- Bailey, E., Bronnenkant, K., Kelley, J., & Hein, G. (1998). Visitor behavior at a constructivist exhibition: Evaluating *Investigate!* at Boston's Museum of Science. In C. Dufresne-Tassé (Ed.), *Évaluation et éducation muséale; nouvelles tendances / Evaluation and Museum Education: New Trends / Evaluacion y educacion museistica: Nuevas tendencias* (pp. 149-168). Montreal: ICOM/CECA.
- Barriault, C., & Pearson, D. (2010). Assessing exhibits for learning in science centers: A practical tool. *Visitor Studies*, 13(1), 90-106. doi: 10.1080/10645571003618824
- Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report*, 13(4), 544-559.
- Bitgood, S., & Gregg, G. (1986). A brief review of the research on signs and labels: Where are the data? *Visitor Behavior*, 1(3), 4-6.
- Bitgood, S., & Patterson, D. (1993). The effects of gallery changes on visitor reading and object viewing time. *Environment and Behavior*, 25(6), 761-781.
- Bitgood, S., & Thompson, D. (1989). The effects of sign length, letter size, and proximity on reading. *Visitor Studies*, 1(1), 101-112.
- Boisvert, D. L., & Slez, B. J. (1995). The relationship between exhibit characteristics and learning-associated behaviors in a science museum discovery space. *Science Education*, 79(5), 503-518.
- Borun, M. (1999). Gender roles in science museum learning. *Visitor Studies Today*, 3(3), 11-14.
- Borun, M., & Dritsas, J. (1997). Developing family-friendly exhibits. *Curator: The Museum Journal*, 40(3), 178-196. doi: 10.1111/j.2151-6952.1997.tb01302.x
- Borun, M., Dritsas, J., Johnson, J. I., Peter, N. E., Wagner, K. F., Fadigan, K., Jangaard, A., Stroup, E., & Wenger, A. (1998). *Family Learning in Museums: The PISEC Perspective*. Philadelphia, PA: Franklin Institute.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavior Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Crowley, K., Callanan, M. A., Tenenbaum, H. R., & Allen, E. (2001). Parents explain more often to boys than to girls during shared scientific thinking. *Psychological Science*, 12(3), 258-261.
- Dancstep, T., Gutwill, J. P., & Sindorf, L. (2015). Comparing the visitor experience at immersive and tabletop exhibits. *Curator: The Museum Journal*, 58(4), 401-422.
- Dancstep, T., & Sindorf, L. (2016). *Exhibit Designs for Girls' Engagement: A Guide to the EDGE Design Attributes*. San Francisco, CA: Exploratorium.

- Dancstep, T., & Sindorf, L. (2018a). Creating a female-responsive design framework for STEM exhibits. *Curator: The Museum Journal*, 61(3), 469-484. doi: doi:10.1111/cura.12268
- Dancstep, T., & Sindorf, L. (2018b). Exhibit Designs for Girls' Engagement (EDGE). *Curator: The Museum Journal*, 61(3), 485-506. doi: doi:10.1111/cura.12267
- Dawson, E. (2014a). Equity in informal science education: developing an access and equity framework for science museums and science centres. *Studies in Science Education*, 50(2), 209-247. doi: 10.1080/03057267.2014.957558
- Dawson, E. (2014b). "Not designed for us": How science museums and science centers socially exclude low-income, minority ethnic groups. *Science Education*, 98(6), 981-1008. doi: 10.1002/sce.21133
- Diamond, J. (1986). The behavior of family groups in science museums. *Curator: The Museum Journal*, 29(2), 139-154.
- Diamond, J. (1994). Sex differences in science museums: A review. *Curator: The Museum Journal*, 37(1), 17-24.
- Falk, J. H. (1983). Time and behavior as predictors of learning. *Science Education*, 67(2), 267-276.
- Falk, J. H. (1997). Testing a museum exhibition design assumption: Effect of explicit labeling of exhibit clusters on visitor concept development. *Science Education*, 81(6), 679-688.
- Falk, J. H., & Dierking, L. (2000). *Learning From Museums: Visitor Experiences and the Making of Meaning*. Lanham, MD: AltaMira Press.
- Feinstein, N. W. (2017). Equity and the meaning of science learning: A defining challenge for science museums. *Science Education*, 101(4), 533-538. doi: doi:10.1002/sce.21287
- Ferguson, C. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538.
- Garibay, C. (2013). *Design Zone Exhibition Summative Report*. Chicago, IL: Garibay Group.
- Girls Inc. (2004). Girls and science, math, and engineering. Retrieved January 19, 2012, from <http://www.girlsinc.org/resources/fact-sheets.html>
- Greenfield, T. A. (1995). Sex differences in science museum exhibit attraction. *Journal of Research in Science Teaching*, 32(9), 925-938.
- Hamilton, L. S., Michael, N. E., Kupermintz, H., Kerkhoven, J. I. M., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: II. NELS:88 Science achievement. *American Educational Research Journal*, 32(3), 555-581.
- Hill, C., Corbett, C., & St. Rose, A. (2010). *Why so Few? Women in Science, Technology, Engineering, and Mathematics*. Washington, D.C.: American Association of University Women.
- Humphrey, T., & Gutwill, J. P. (Eds.). (2005). *Fostering Active Prolonged Engagement: The Art of Creating APE Exhibits*. Walnut Creek: Left Coast Press.
- Koke, J. (2007). [Expert interview].
- Kremer, K., & Mullins, G. (1992). Children's gender behavior at science museum exhibits. *Curator: The Museum Journal*, 35(1), 39-48.
- Lee, V. E., & Burkam, D. T. (1996). Gender differences in middle grade science achievement: Subject domain, ability level, and course emphasis. *Science Education*, 80(6), 613-650.
- National Science Foundation. (2003). *New Formulas for America's Workforce: Girls in Science and Engineering*. Arlington, VA: NSF 03-207.
- National Science Foundation. (2007). *New Formulas for America's Workforce 2: Girls in Science and Engineering*. Arlington, VA: National Science Foundation.
- Olszewski, B., Macey, D., & Lindstrom, L. (2006). The practical work of < coding>: An ethnomethodological inquiry. *Human Studies*, 29(3), 363-380.

- Oppenheimer, F. (1986). *Working Prototypes: Exhibit Design at the Exploratorium*. San Francisco, CA: Exploratorium.
- Perry, D. (1989). *The Creation and Verification of a Developmental Model for the Design of a Museum Exhibit*. Doctoral Dissertation, Indiana University.
- Perry, D. (1993). Experiment Bench Project: Summative Evaluation Report. Saint Paul, MN: Science Museum of Minnesota.
- Perry, D. (2012). *What Makes Learning Fun? Principles for the Design of Intrinsically Motivating Museum Exhibits*. Lanham, MD: AltaMira Press.
- Peterson, P., & Fennema, E. (1985). Effective teaching, student engagement in classroom activities, and sex-related differences in learning mathematics. *American Educational Research Journal*, 22(3), 309-335.
- Reich, C. (2005). *Universal Design of Interactives for Museum Exhibitions*. Master of Education, Lesley University, Cambridge.
- Rosser, S. V. (1991). *Female Friendly Science: Applying Women's Studies Methods and Theories to Attract Students*. New York: Teacher's College Press.
- Sammet, K., & Kekelis, L. (2016). *Changing the Game for Girls in STEM: Findings on High Impact Programs and System-Building Strategies*.
- Sandifer, C. (2003). Technological novelty and open-endedness: Two characteristics of interactive exhibits that contribute to the holding of visitor attention in a science museum. *Journal of Research in Science Teaching*, 40(2), 121-137.
- Sanford, C. W. (2010). Evaluating family interactions to inform exhibit design: Comparing three different learning behaviors in a museum setting. *Visitor Studies*, 13(1), 67 - 89.
- Screven, C. (1995). Motivating visitors to read labels. In A. Blais (Ed.), *Text in the Exhibition Medium* (pp. 97-132). Quebec: Musee de la Civilisation.
- Serrell, B. (1996). *Exhibit Labels: An Interpretive Approach*. Walnut Creek, CA: Alta Mira Press.
- Serrell, B. (1998). *Paying Attention: Visitors and Museum Exhibitions*. Washington, D.C.: American Association of Museums.
- Serrell, B. (2010). Paying more attention to paying attention. *Center for Advancement of Informal Science Education*.
- Serrell, B. (2015). *Exhibit Labels: An Interpretive Approach* (2<sup>nd</sup> ed.). London, UK: Rowman & Littlefield.
- Sindorf, L., & Dancstep, T. (2015). Exhibit Designs for Girls' Engagement (EDGE) Exhibit Coding Scheme and Protocol. San Francisco, CA: Exploratorium
- Taylor, D. (2005). Social science: Observing women and girls in the museum. *ASTC Dimensions*(May/June), 11-12.
- Van Schijndel, T. J. P., Franse, R. K., & Raijmakers, M. E. J. (2010). The Exploratory Behavior Scale: Assessing young visitors' hands-on behavior in science museums. *Science Education*, 94(5), 794-809. doi: [10.1002/sce.20394](https://doi.org/10.1002/sce.20394)
- Verheyden, P. (2003). The great sexperiment. *ECSITE Newsletter*, 54(Spring), 10-11.
- Wellcome Trust. (2013). *Risks and Rewards: How PhD Students Choose Their Careers: Qualitative Research Report*. London, UK.
- Wilson, C., Voorhis, V., & Morgan, B. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2), 43-50.
- Yalowitz, S. S., & Bronnenkant, K. (2009). Timing and tracking: Unlocking visitor behavior. *Visitor Studies*, 12(1), 47-64. doi: [10.1080/10645570902769134](https://doi.org/10.1080/10645570902769134)