Title: Exhibit Designs for Girls’ Engagement (EDGE)

Introduction

Informal science centers are an important resource in the effort to increase STEM interest and engagement (Falk and Dierking 2010, National Research Council 2009, National Science Board 2008, Salmi 2001, 2002). By offering hands-on interactivity and playful exploration, studies show museums can foster scientific thinking and engagement—key aspects of participation in science (Hamilton et al. 1995, Humphrey and Gutwill 2005, National Research Council 2009, Randol 2005). Others have found that people with science careers often credit their initial interest in STEM to memorable museum visits (Cosmos Corporation 1998, McCreedy and Dierking 2013, Salmi 2002).

Yet, research has shown that girls are less likely to visit science museums than boys (Borun 1999, Hamilton et al. 1995, National Science Foundation 2003). Within the walls of science museums, researchers have found that girls are less likely to use exhibits about certain content areas, such as physics and engineering (Diamond 1994, Girls Inc. 2004, Greenfield 1995, Kremer and Mullins 1992, National Science Foundation 2003, Ramey - Gassert 1996, Verheyden 2003).
Similarly, research has indicated that many girls spend less time engaging at STEM exhibits (Diamond 1994, Girls Inc. 2004, Greenfield 1995, Kremer and Mullins 1992, National Science Foundation 2003, 2007, Ramey-Gassert 1996, Verheyden 2003). These findings suggest that science museums must consider whether the experiences they are offering could be designed to better engage girls.

Science museums and researchers worldwide are in fact working to better engage girls in STEM topics (Achiam and Holmegaard 2015, Brown, Huerta Migus, and Williams 2012, Cardella, Svarovsky, and Dorie 2013, Chatman et al. 2008, Laursen 2011, Munley and Rossiter 2013, National Science Foundation and National Center for Science and Engineering Statistics 2013, Ramey-Gassert 1996, Roughneen 2011). Some science museums offer special programs for girls and women, and a few are creating exhibitions highlighting women in STEM or exploring gender differences and inequities (Laursen 2011, National Science Foundation 2003, 2007, Verheyden 2003). Best practices from the field suggest science museums may also improve exhibit explorations and confidence when they offer experiences that align with girls’ learning preferences (Baker 2013, Munley and Rossiter 2013).

Despite these efforts, few research studies have explored ways of better engaging girls in a free-choice learning environment. Those that have were focused on a single STEM topic, design feature, or interactive exhibit (Cardella, Svarovsky, and Dorie 2013, Crowley et al. 2001, Dancu 2010, Sinkey, Rosino, and Francisco 2014) or best practices rather than rigorous research (National Science Foundation 2003, 2007). The study presented in this paper builds on those prior findings, exploring the potential of a female-responsive framework (described below), by considerably expanding the scope of study to include the engagement of over 900 children across three institutions at more than 300 STEM exhibits representing over 55 design attributes.
The Female-Responsive Design Framework

The NSF-funded study described in this paper aimed to identify the most important exhibit design attributes for engaging girls aged 8-13. The first step in this process was the development of our Female-Responsive Design (FRD) Framework (described in detail in Dancstep and Sindorf 2018 this issue). Using the theoretical lens of Culturally Responsive Pedagogy (CRP), the framework defines “female” inclusively, acknowledging that gender occurs along a spectrum and that there is considerable variability among females. At the same time, the CRP model enables the framework to identify patterns across females’ social, historical, and cultural repertoires of STEM learning, in order to better support females’ engagement. The framework synthesizes research from the fields of psychology, education, museum learning, gaming, and web design—leveraging existing qualitative and quantitative work, as well as interviews with practitioners and experts. The FRD Framework describes four pedagogical strategies to engage and support females’ STEM learning: Enable Social Interaction and Collaboration; Create a Low-Pressure Setting; Provide Meaningful Connections; and Represent Females and their Interests.

Within these four strategies, the FRD Framework identifies 55 exhibit-specific female-responsive design attributes that have the potential to better engage females. Table I provides examples of our approach to identifying specific exhibit design attributes that fit within the broad strategies of the FRD Framework. (For the full list of strategies, attributes, and related references, see Dancstep and Sindorf 2018 in this issue). These design attributes were developed in concert with expert advisors and practitioners, along with precious few research studies. The next phase of our research—the current study—presented an opportunity for a deeper investigation of the design attributes to determine which ones best engaged girls at physics, engineering, math, and perception exhibits.

[Place Table I about here. Table I. Sample Female-Responsive Design Strategy and Adaptation to Exhibit Design.]

Current Study

This study—Exhibit Designs for Girls’ Engagement (EDGE)—sought to learn:

This article is protected by copyright. All rights reserved
Which of the many design attributes identified in the FRD Framework are most strongly related to girls’ engagement at physics, engineering, math, and perception exhibits?

We chose an exploratory approach for this study because it is the first attempt at this scale to explore the relationship between exhibit designs and girls’ engagement. While much of the research referenced in the FRD Framework drew on small case studies (such as those assembled by the National Science Foundation 2003, 2007), this stage of the research trajectory is important because it sets us up to investigate those recommendations with a broader scope. The study entailed investigating a large number of design attributes (n = 55, plus variations) based on the engagement of numerous girls (n = 450) at a variety of exhibits (n = 334) across multiple institutions (n = 3). To date, a study of this size has not been implemented in the realm of exhibit design. Such research and design studies typically involve a smaller number of exhibits and design attributes. Previous studies of exhibit designs have ranged from single exhibits to exhibition collections of 30-60 exhibits, and typically focus on only a handful of design attributes (see for example, Borun and Dritsas 1997, Danecstep, Gutwill, and Sindorf 2015, Gutwill-Wise and Allen 2002, Perry 2012, Sandifer 2003, Sanford 2010). The large number of design attributes identified in the FRD Framework necessitated a much broader investigation. Examining the relationship between female-responsive exhibit design attributes and girls’ engagement at a range of STEM exhibits allowed us to narrow the large set of promising design attributes in the FRD Framework and create a smaller more practitioner-friendly set based on girls’ actions and interactions with exhibits.

This paper describes our large-scale quantitative investigation, which identifies the most promising of the design attributes for girls. However, this investigation is just one part of a larger research project that employed an exploratory sequential mixed-methods approach. Mixed-methods investigations have been employed by other educational and sociocultural researchers seeking to understand complex phenomena (e.g., Creswell 2009, Crook and Garratt 2011, Gay 2010, Mercer et al. 2004, Ogbu and Simons 1994, Scribner and Cole 1981, Vann and Cole 2004). In a separate article, we report a qualitative study which triangulates the quantitative findings identified here (Dancstep and Garcia-Luis Forthcoming).
Methods

The EDGE team collected and correlated two sets of data. The first aimed to assess exhibit design attributes at a large variety of interactive STEM exhibits, and the second aimed to understand girls’ (and boys’) engagement at those exhibits.

Exhibit Design Attributes: Exhibits and Procedure

The EDGE study included 334 interactive exhibits. This extensive approach allowed us to gather information on a variety of exhibit types and assess our comprehensive list of potentially engaging exhibit design attributes. We studied exhibits at three U.S. institutions, chosen to expand the generalizability of our findings. The three institutions varied in size, region, demographic representation, approach to exhibitry, and number of focal exhibits available to study. The research focused on exhibit topics where broader gender gaps in interest and participation persist, including physics, math, and engineering, rather than biology or psychology (National Science Foundation and National Center for Science and Engineering Statistics 2013). In the end, we studied 213 exhibits at a West Coast museum, 70 at a Midwest museum, and 18 at a Southwest museum. To address differences in the number of exhibits represented at each institution, we used analytic approaches, such as exploring the results on an institutional basis and controlling for institution (details below). During analysis, 33 exhibits were dropped from the study because they were broken, removed from the study area, or had too few users to analyze. The final study included 301 exhibits. Figure A details the attrition of exhibits from our sample.

Exhibit Design Attributes: Measurement

The EDGE team developed a tool to assess the presence or absence of each of the 55 design attributes (Appendix B provides the full list of attributes). Drawing from the literature whenever possible, we developed one hundred questions about observable exhibit attributes. For example, *The exhibit includes more than one station: Yes or No.* For some attributes the tool employed a scale to evaluate the degree to which the attribute was present, and then in the analysis phase the attribute was considered “present” if it was above a certain threshold. The tool included assessment of variations, such as whether an attribute was noticeable from afar or up close, or
whether an attribute was central or peripheral to the exhibit experience. The tool also included photos of example exhibits with and without each of the design attributes. All items were examined and adjusted for face and content validity via discussions with advisory groups and practitioners.

Two researchers, blind to the hypotheses of the study, independently scored all exhibits for the design attributes. The pairs then met to compare their scores, discussing and determining final codes for each attribute at each exhibit. This process of double coding followed by resolution of disagreements was used to improve reliability (Baxter and Jack 2008, Olszewski, Macey, and Lindstrom 2007).

Two of the design attributes presented a potential challenge for our adult researchers to assess, because they considered girls’ personal experiences: whether the exhibit contained a familiar object, and whether the exhibit was relevant to girls’ lives. As with the other design attributes, we developed exhibit observation questions for these attributes. As a small complementary check, researchers also approached three girls at each exhibit, asking two closed-ended interview questions regarding familiarity and relevance.

To include a design attribute in our analysis, we required a sample of 50 exhibits with that attribute. A minimum sample size of 50 exhibits is large; however we wanted to have high confidence in our results and chose this number based on the minimum cell size of 50 for correlations and regressions (Wilson, Van Voorhis, and Morgan 2007). Thus, we removed design attributes found at fewer than 50 exhibits (detailed in Appendix B). This left us with a remaining 34 potential design attributes and their variations.

**Control Variables: Institution and Crowdedness Tendency**

After exploring several possible alternative explanations for the relationships between design attributes and girls’ engagement, we identified institutional differences and crowdedness as important control variables. Noting an exhibit’s institution helps account for differences such as museum size, pedagogical approach, visitor demographics, location, and the number of exhibits studied. Crowding has been identified as a limitation to studying exhibit engagement (Sandifer
2003); thus, we created a Crowdedness Tendency score for each exhibit. To create the Crowdedness Tendency score, we first assessed the average number of visitors typically at an exhibit, by photographing each exhibit and counting the number of visitors present. We repeated this at least six times for each exhibit, at varying times (mid-morning and afternoon) and days of the week (one weekday and one weekend day) throughout the tracking and timing data collection. Museum staff also estimated the number of users the exhibit appeared to be designed for. The Crowdedness Tendency score was created by dividing the average count of visitors at each exhibit by the estimated potential users.

**Girls’ Engagement: Participants and Procedures**

To gather engagement data, we randomly selected children attending a museum with their families, and invited them to participate in the study. In order to maintain equal sample sizes throughout data collection, we oversampled within each institution whenever imbalances greater than five occurred between girls and boys, younger and older children, or recruit locations on a museum’s floor. We studied 906 participants, including 637 from the West Coast museum, 131 from the Midwest museum, and 138 from the Southwest museum. Because visitors only use a portion of exhibits in a space (Serrell 1998), we included a larger number of participants at the West Coast museum, which offered a greater number of exhibits for study. Beyond that, the discrepancy in the number of participants at each institution does not affect the study results: the number of visitors only affects the precision of our average engagement measures on a per exhibit level. For that reason, any exhibits with fewer than ten visitors were dropped. Further, the analyses combine all data across exhibits and their participants to assess the design attributes; our analytic approach required at least 50 exhibits represent each design attribute, ensuring high precision of our engagement measures for each design attribute. See Table II for participant demographics.

While our study focused on girls, we collected data on all genders, so that we could pursue an inclusive approach, and so we could investigate and discuss any negative effects on boys’ engagement. Participating children chose their own gender from a dropdown list that included female, male, and a custom text box option. Participants also reported their age; this study focused on children ages 8-13. We targeted this age group because research shows that

This article is protected by copyright. All rights reserved
elementary and middle school is an especially important time to spark interest in STEM (President’s Council of Advisors on Science and Technology 2010). This is the age when interest in STEM begins to decline, and at a faster rate for girls (Bailey et al. 1992, Miller, Slawinski Blessing, and Schwartz 2006). The team strove for an ethnically diverse visitor sample by including three museums in the study, each in a city with different demographic representation. We also increased data collection efforts on days when audiences were particularly diverse, such as free days and Latino family program days, which are aimed at underrepresented cultural and economic communities.

[Place Table II about here. Table II. Participant demographics.]

**Girls’ Engagement: Measurement**

The project employed traditional timing and tracking methodology for measuring children’s engagement at all exhibits in the study (Serrell 1998, 2010, Yalowitz and Bronnenkant 2009, Yalowitz and Ferguson 2006). Tracking and timing allows for unobtrusive observation of participants as they move through a museum space, and is a common methodology used for understanding visitors’ engagement and behaviors in museum settings (Serrell 1998, 2010). Different exhibit design attributes may afford different types of engagement; therefore, we chose to explore four separate measures of engagement: Use, Return Visits, Time Spent, and Engagement Level scores. Each of these measures informs us about interesting yet separate aspects of the exhibit experience (Sanford 2010). (For some measures, it was necessary to remove a few exhibits from the analysis; details are provided below and Figure A summarizes exhibit attrition.) Fifteen percent of participants were inter-rated, and all researchers collecting data were blind to the study’s purpose.

**Use:** A common measure of attraction power used in free-choice settings is the percentage of visitors who stop to use an exhibit (Diamond 1986, Falk 1983, Humphrey and Gutwill 2005, Sandifer 2003, Serrell 1998). Because we employed a tracking and timing methodology, we calculated the percentage of tracked children who used each exhibit. We considered participants to have used an exhibit if they touched it for three or more seconds; or stopped, feet planted,
facing the exhibit for five or more seconds. Inter-rater agreement was high at 93% for all instances of Use.

To minimize any confounding effects of exhibit location, we removed any exhibits that were in low-traffic areas. In the end we had a final sample size of 269 exhibits for the Use analyses.

**Return Visits:** We posit that return visits are also indicative of engagement, though used less commonly (Danestep, Gutwill, and Sindorf 2015). Similar to other indictors of engagement in situations where learners are free to choose, we believe returning to use an exhibit again suggests higher interest and engagement (Diamond 1986, Falk 1983, Humphrey and Gutwill 2005, Sandifer 2003, Serrell 1998, Yalowitz and Bronnenkant 2009). We measured 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. Again, inter-rater agreement for all instances of Use, including Return Visits, was high at 93%. The final exhibit sample size for Return Visits was 301 exhibits.

**Time Spent:** A common measure of holding power in free-choice settings is the average time visitors spend at an exhibit (Diamond 1986, Falk 1983, Humphrey and Gutwill 2005, Sandifer 2003, Serrell 1998, Yalowitz and Bronnenkant 2009). We measured the average time participants spent at each exhibit. Time Spent sums all instances of Use, from when a child first uses an exhibit (see above) to when the child walks away, or faces away from the exhibit for 5 or more seconds. Inter-rater agreement for Time Spent across all instances of exhibit use was high with Spearmann’s Rho at .94 (Ferguson 2009).

While some exhibits are designed for open-ended, lengthy use, others intentionally offer a quick-hit experience. In the latter instances, Time Spent is not a good measure of engagement. Therefore, we identified exhibits where the average Time Spent was less than 30 seconds, and removed those where researchers and practitioners judged that the exhibits were designed to offer a short experience (Humphrey and Gutwill 2005). The final exhibit sample size for Time Spent was 279 exhibits.
Engagement Level (EL): Observing visitors’ various interactive behaviors to better understand their engagement and learning is also common practice in the museum setting (Barriault and Pearson 2010, Borun and Dritsas 1997, Humphrey and Gutwill 2005, Sanford 2010, Van Schijndel, Franse, and Raijmakers 2010). We drew on two existing metrics (Barriault and Pearson 2010, Van Schijndel, Franse, and Raijmakers 2010) to create a scale that would give us a sense of the depth of children’s engagement at exhibits. These prior measures had been developed for use with a small number of exhibits (six or fewer), and with reference to exhibit-specific criteria; we adapted the scales to address the larger variability introduced by greatly expanding the number of exhibits under study (i.e., from six to 300).

The amended Engagement Level (EL) rubric (Table III) was piloted with five exhibits chosen for their design differences, to ensure variability in children’s use patterns. For the full study, we recorded the Engagement Level each participant reached at each exhibit. However, the measure had less variability when moving from ~20 children at five exhibits in the pilot phase to over 900 children at 301 exhibits in the full data set. The majority of children reached the highest Engagement Level, Level 3 (ranging from 70% - 100% of visitors). Given the interactive nature of science museums, this may not be surprising; yet, such range restriction makes our measure less sensitive to correlations, meaning that some of the relationships between the design attributes and the Engagement Level measure may be underestimated (i.e., lower than we would find with less range restriction). With the majority of visitors reaching Level 3 we determined that the most useful unit of analysis would be whether or not they reached the highest engagement level. From here on, this measure will be referred to as the Highest Engagement Level. To average across visitors, that score is represented by the percentage of girls who reached the Highest Engagement Level at each exhibit.

[Place Table III about here. Table III. Engagement Level rubric.]

Inter-rater agreement was high at 92%. However, we wanted to use a more stringent measure of reliability, Kappa, to control for chance agreement. This approach resulted in reliability that varied widely by exhibit. The following explains the challenges that led to this variability and describes how we’ve adjusted our approach to maintain high confidence in the results.
Again, this is an extremely large data set, and with that came a few complications. While the large range of kappa statistics across exhibits could indicate a lack of sensitivity in the assessment for certain kinds of exhibits or designs, the high percentage agreement suggests that the measure was applicable to most exhibits. Ultimately, we found that the main reason the reliability varied per exhibit was due to the fact that reliability was captured for 15% of participants, which often led to a very low percentage of inter-rated instances for a particular exhibit, and therefore, low inter-rater scores if any disagreements occurred. This issue was greatest at the Midwest museum with a large number of exhibits, n = 70, but a lower number of participants, and therefore a low number of inter-rated participants, n = 17. Further, at the Midwest museum, one data collector became unblinded during data collection; checks for any potential bias revealed that that data collector’s EL scores were, at times, significantly different from the other data collectors, and that same data collector was overrepresented in the reliability data collection. Ultimately, we decided to remove the Midwest museum exhibits from the Highest Engagement Level analyses, and to create an inter-rater reliability threshold that each exhibit needed to reach to be included in the analyses. This smaller sample size (as well as the exclusion of one of the institutions) may limit the generalizability of the results; however, the number of exhibits and institutions is still larger than prior studies of engagement and exhibit design, allowing for confidence in the findings. This approach resulted in a final exhibit sample size of 81 for the Highest Engagement Level analyses, with an average Kappa of .72 (which Landis and Koch 1977 consider substantial).

[Place Figure A about here. Figure A. Exhibit Attrition. This figure details the reasons for removing some exhibits from certain analyses, and gives the final number of exhibits included or excluded from each.]

**Results**

The EDGE project was always intended to be useful for practitioners. The results of this investigation generated an extraordinarily large set of data that is uncommon in exhibit design research. This presented two problems for the project, one practical and one analytical. On the practical side, we knew it would be difficult for exhibit teams to consider the FRD Framework’s
full set of design attributes (Dancstep and Sindorf 2018 this issue) while developing an exhibit. Therefore, we wanted to reduce the number of attributes by identifying the most promising subset for designing with young girls in mind\(^2\). To do that work, we examined the attributes to determine which were the most strongly and positively related to girls’ engagement at STEM exhibits. After obtaining the results based on girls’ data, we returned to the boys’ data to check for any ill effects on boys’ engagement.

Analytically, the large number of design attributes, along with the comprehensive set of engagement measures, also presented a statistical problem: conducting too many inferential analyses can increase the likelihood of finding a significant relationship simply by chance. To mitigate that possibility, we began our analysis process by using descriptive statistics to reduce the number of inferential analyses necessary. Below we describe our analysis of girls’ data in more detail, summarized in Figure B.

We began identifying the most promising attributes by using correlational data to explore the relationships between each design attribute (and its variations) with each of the engagement measures. (Note that correlations are descriptive until inferential p-values or confidence intervals are applied, which was not done during this step.) We were looking for the attributes that were positively related to girls’ engagement—that is, present when girls were highly engaged and absent when girls were less engaged. Therefore, we eliminated any attributes that had a significant negative relationship with any engagement measure at any institution—where the presence of a design attribute resulted in lower engagement for girls, even if the relationship was positive for other measures or institutions. Looking at the relationships on an institutional basis helped ensure that the findings did not stem from stronger weightings for the institutions with more exhibits in the dataset. Next, we narrowed the set by considering attributes or their variations only if they were strongly positively correlated with the following:

- at least two engagement measures at one institution; or
• at least one engagement measure at two institutions; or
• one engagement measure across all three institutions combined, as long as the result was not being driven by a single institution.

From the remaining set of design attributes, we further reduced the set by identifying those that were ultimately redundant with one another. Redundant attributes included:
• subtle variations (the attribute is visible from up close versus the attribute is visible from afar; the attribute is central to the exhibit phenomenon versus the attribute is peripheral to the exhibit phenomenon);
• definitional variations (The exhibit has bright, prominent color versus The exhibit color palette is vivid); and
• certain combinations (Can be used from multiple sides and Has multiple stations versus The exhibit has multiple stations or sides).

Design Attributes were considered to be redundant only if their relationships to each of the engagement measures were similar. For each group of redundant attributes, the specific instantiation with the strongest correlations was kept in the analysis while the others were eliminated as redundant.

Through these correlational analyses we identified a subset of eleven specific exhibit design attributes that were strongly positively related to girls’ engagement across multiple measures or institutions. However, we wanted to ensure that those relationships could not be explained by other features of the museum environment.

Checking for Alternative Explanations
We conducted a separate regression analysis for the 11 remaining design attributes and each engagement measure in order to control for two important alternative explanations: Institution (such size, visitor makeup, and number of exhibits studied) and Crowdedness Tendency. We chose to run the regressions separately because we were simply checking to make sure the relationship between each of the design attributes and girls’ engagement was not due to these alternative explanations. Keeping the design attributes separate is consistent with the flexibility recommended by Culturally Responsive Pedagogy. Further, this approach allowed us to focus on results that could best support the way museum professionals would apply them in practice: picking among a set of options based on the exhibit and its needs, rather than implementing the
full suite of attributes. Running each analysis separately, while not a common practice, has been used in other fields (Crain et al. 2014).

Of the 11 design attributes included in the regression analyses, nine were significantly related to at least one engagement measure across the three science centers, while controlling for institutional differences and crowdedness. Tables IV and V provide the results of the final analyses—reporting significance, interpreting the slope (or Beta), and supplying the effect sizes.

Appendix A gives the results using more traditional output. Note that the strength of the results depends on which engagement measure or combination of engagement measures are under consideration. In practice, we hope educators will choose among the nine EDGE Design Attributes based on which of the attributes makes the most sense for a particular exhibit or set of exhibits rather than solely on the strength of the results across the four engagement measures.

When assessing the magnitude of these findings, it is best to consider the effect sizes and the regression slope (or Beta). For effect sizes, we report the semi-partial correlations \(sr\), which provide the unique relationship between each design attribute and each engagement measure, over and above crowdedness and institution. Semi-partial correlations are considered small but meaningful if they are greater than .10, moderate if they are around .3, and large if around .5 or greater (Cohen 1988, 1992). The small effect sizes are meaningful, especially given the context: for example, in cases where we find that 5% more girls use exhibits when the attribute is present, that result could have a large effect over the thousands—or hundreds of thousands—of visitors to such institutions. For example, at a single institution, such as the West Coast museum, that has around 800,000 visitors per year, of which 7% are girls aged 8-13, this could add up to almost 3,000 more girls per year using exhibits which have those design attributes. Further, the interactive exhibits in our study had average holding times of approximately 60 seconds for girls, so increasing girls’ time spent by eight or more seconds could be a big impact in this situation.

The Final EDGE Design Attributes

Table IV presents the resulting EDGE Design Attributes along with the strength of their relationship to each engagement measure. To sort our table in a meaningful way, we’ve ranked...
each Design Attribute per engagement measure, calculated the mean of the four rankings, and sorted by the mean rank. This places the attributes that have the strongest relationships to each of the four measures of engagement at the top, while not making comparative claims.

[Place Table IV about here. Table IV. The final EDGE Design Attributes.]

For each of the nine EDGE Design Attributes listed in Table IV, we examined data from 450 boys and found that none were harmful to boys’ engagement across any of the measures. In some instances, but not in as many as there were for girls, the attributes were also useful for boys’ engagement.

From the subset of 11 design attributes subjected to regression analyses, two were no longer meaningfully related to girls’ engagement once we controlled for institutional differences and crowdedness. These results are detailed in Table V.

[Place Table V about here. Table V. The design attributes that were no longer significant after controlling for institution and crowdedness.]

By exploring the relationship between girls’ engagement and design attributes across an extensive number of participants and exhibits, we were able to identify which of the many design attributes identified in the FRD Framework are most promising for engaging girls at STEM exhibits. From the full list of 55 design attributes, several were not strongly positively related to girls’ engagement, and others were actually negatively related to their engagement in some instances (see Appendix B for the ultimate outcome for each of the design attributes). Next we discuss the overall results.

Discussion

This exploration allowed us to test the applicability of the FRD Framework as a tool for identifying exhibit designs that better engage females. The large-scale and comprehensive study
allowed us to identify the best design attributes for engaging girls ages 8-13 in the context of physics, engineering, math, and perception exhibits. We know that each exhibit affords different design attributes and we encourage project teams to choose the most appropriate among the nine EDGE Design Attributes based on their own expertise and the specific needs of an exhibit or set of exhibits. Further details about assessing and implementing the EDGE Design Attributes can be found in the practitioner’s guide (Dancstep and Sindorf 2016). A museum experience that incorporates exhibits with a mix of these attributes holds much potential for appealing to a variety of girls.

Below we discuss the nine most promising design attributes along with some of the attributes that did not rise to the top. Throughout, we provide quotes from the Girl Advisory Committee (GAC) to help illuminate the results.

Our findings are consistent with prior work regarding females’ patterns of engagement with open-ended and playful, whimsical designs. Some accounts have posited that open-ended, playful activities, such as those found in maker spaces, reduce pressure by widening definitions of intelligence and creating opportunities for success (Vossoughi et al. 2013). Recent work has also identified open-ended and playful, whimsical STEM experiences as especially appealing to females’ interests (e.g., Ford et al. 2006, Haste 2004, Intel Harris Poll 2014, Jenkins 2001). Several GAC advisors talked about the benefit of open-ended exhibits, explaining, for example, “If there’s no way to get it wrong, you can just have fun with it.” Moreover, one GAC advisor told us why she liked whimsical exhibits: “There’s a lot of moments in my life that are serious, where you have to take things seriously, like school and stuff. It’s fun to just relax, play a little, so life’s not so serious.”

Design aesthetics may impact girls’ feelings of belonging and comfort in a science center environment. Researchers have found that girls often view STEM as stereotypically masculine (Archer et al. 2012, 2013), and while these stereotypes may be activated or perpetuated through design cues (Cheryan et al. 2009, Master, Cheryan, and Meltzoff 2016), they may also be mitigated by them. In addition to whimsical design, homey or personal design may help mitigate “geeky” or masculine stereotypes of science (Master, Cheryan, and Meltzoff 2016).
Similarly, **homey aesthetics** may be related to feelings of safety and comfort, which Gontan (2013) found to be important for engaging Latina girls in STEM programs. One of our girl advisors raised the idea of comfort when describing the appeal of the homey exhibit aesthetic: “It feels like you’re at home, like you can just relax.”

Researchers have reported that females often have low confidence in their STEM abilities (e.g., Chatman et al. 2008, Nix, Perez-Felkner, and Thomas 2015). Our Girl Advisory Committee members echoed this sentiment throughout the project. For example, one GAC advisor told us, “I didn’t want to do one of the complicated ones. I was afraid I’d mess up, so I only wanted to do an easy one.” Providing illustrative **use drawings** and **opportunities to watch others** allows for a preview of the experience, helping to orient girls and potentially boosting their comfort with the activity (Perry 2012, Rosser 1991). Our study adds to findings from other researchers noting that when museum **visitors watch others** using an exhibit, they may decide to use an exhibit, become more engaged with it, or learn what to do or notice (Heath et al. 2002, Meisner et al. 2007, vom Lehn, Heath, and Hindmarsh 2001). In a similar vein, including a **use drawing** may help girls feel comfortable getting started, or encourage them to persevere as they use the exhibit (Fenichel and Schweingruber 2010, Perry 2012, Serrell 2015).

Our findings align with prior research regarding females’ inclinations toward social and collaborative learning environments (Diamond 1994, Humphrey and Gutwill 2005, National Science Foundation 2003, Peterson and Fennema 1985, Rosser 1991, Taylor 2005). Exhibits with **space to accommodate three or more people** may help keep groups together, allowing for a more social experience. As a GAC advisor explained: “I came with ALL of my friends…I wanted to see the museum WITH them.” Along those same lines, exhibits with **multiple sides or stations** and those with **open-ended designs** have been shown to encourage social interaction and inquiry-oriented conversation (Borun and Dritsas 1997, Humphrey and Gutwill 2005). A member of the Girl Advisory Committee described why collaborative exhibit designs appealed to her: “With someone else, they have the same problems as you. You figure it out with them. You don’t feel alone; you have help.”
Finally, our findings are consistent with prior reasoning that when science is integrated with everyday activities, objects, and materials, it can help make science content feel more relevant and familiar, thereby bolstering girls’ interest and self-confidence (Baranowski and Delorey 2007, Ford et al. 2006, Froschl et al. 2003, Jones, Howe, and Rua 2000, Kekelis, Heber, and Countryman 2005, Koke 2005, Maher 2005, McCreedy 2005, Resnick, Berg, and Eisenberg 2000, Rosser 1991, Taylor 2005, Vossoughi and Bevan 2014). Girls’ higher engagement at exhibits with images of people, familiar objects, and homey aesthetics supports the notion of using meaningful connections to pique girls’ interest and make science feel more approachable.

As previously mentioned, the EDGE project was intended to be useful for practitioners, and we aimed to reduce the number of attributes practitioners would need to keep in mind. To that end, we removed any attributes that were not universally positively related to girls’ engagement. Below we explore some of the reasons that design attributes may not have made it into the final, most promising, set. (Appendix B identifies the final outcome for the full set of design attributes.)

First, some of the dropped design attributes support the overall approach of the Female-Responsive Design Framework. Grounded in Culturally Responsive Pedagogy, the framework asserts that educators should attend to the specifics, nuances, and variability in the ways pedagogical approaches manifest in different contexts. The FRD Framework includes four broad strategies, but recommends considering the nuanced design approaches within each strategy rather than assuming that any approach within the broad strategies will be effective. For example, while the broad strategies might suggest that any opportunity for social interaction will benefit girls, we found that some design attributes having to do with social interaction, such as seating for multiple people or designed for multiple players to use without interfering with each other, were negatively related to some measures of girls’ engagement at one or more institutions (while other social attributes remained positive and important). These differences may depend on a particular museum’s context, or the age group or topic of interest; such findings remind us that it is important for the field to continue exploring the larger set of design attributes within the FRD Framework.
Next, the FRD Framework drew on research from a variety of contexts, but design attributes that worked well in their initial domain did not necessarily extend to a museum environment. For example, work in the field of web design has suggested that women have a preference for many bright colors (Moss, Gunn, and Heller 2006, Moss, Gunn, and Kubacki 2007); however, we found that exhibit designs with *bright, prominent color* did not ultimately show a strong relationship with girls’ engagement after controlling for institutional differences and exhibit crowdedness. Another example: *label describes how the exhibit phenomenon is used in the real world* was negatively related to some measures of girls’ engagement. Certain strategies, such as connecting to real-world applications, may be better implemented during classroom or afterschool activities (Kekelis, Heber, and Countryman 2005, Maher 2005), rather than at an exhibit. These findings warrant further exploration. In the meantime, when implementing the less supported design attributes, practitioners will want to pay especially close attention to their own contexts and communities.

Finally, we were unable to test some of the promising attributes because too few exhibits in our study offered those designs. In two cases—*exhibit has embedded story or narrative* and *exhibit topic has been related to shared male and female interests via prior research or evaluation* —the supporting research is strong (see e.g., Casey, Andrews, et al. 2008, Casey, Erkut, et al. 2008, Ford et al. 2006, Osborne and Dillon 2008, Sjøberg and Schreiner 2010); in another, *self-image*, the attribute was proposed by our Girl Advisory Committee. The aforementioned attributes, along with *exhibit provides two or more roles, label invites visitors to work together, and exhibit involves self-expression or authorship*, all had strong positive relationships to girls’ engagement. We feel that studying these attributes will be the most important next step for understanding the best design attributes for engaging girls at STEM exhibits.

**Conclusion**

The Exhibit Designs for Girls’ Engagement (EDGE) study has begun the work of identifying promising design strategies to better engage girls at STEM exhibits. The EDGE Design Attributes help to expand the definition of “good exhibit design” to incorporate female-inclusive practices (Feinstein 2017). Our broad, large-scale approach allowed us to establish results that generalize across a variety of exhibit types. We posit that integrating these attributes into exhibit
experiences enhances museums’ potential for appealing to a multitude of girls; we encourage museum educators to add the nine EDGE Design Attributes to their practice whenever possible.

There are many promising avenues for continuing the Informal Science Education field’s efforts to be more female-inclusive. For example, beyond the correlational research design we employed, it is important to explore how each of the EDGE Design Attributes might directly, comparatively, or jointly impact girls’ engagement when applied to the same exhibit. Further, the field must learn how the Design Attributes impact the museum experience when considering intersections of gender, cultural, and STEM identities, and the ways these shift over time (Gay 2002, 2010, Modi, Schoenberg, and Salmond 2012, Paris 2012, West and Fenstermaker 1995). In the end, exhibit design teams should embrace a culturally responsive approach to pedagogy (CRP), acknowledging the diversity among girls, and tailoring their designs to fit the nuanced needs of their particular contexts and communities. We look forward to the ongoing conversation as the ISE field continues to work toward greater gender inclusivity in STEM.

[Place Appendix A in appendix. Appendix A. Full regression analysis results.]

[Place Appendix B in appendix. Appendix B. Full list of design attributes and their results.]

Endnotes

1. We relied on random selection; however, if we detected a large difference in the numbers of boys and girls recruited, we switched to purposive sampling until we balanced the samples.

2. We considered clustering the design attributes into a smaller set of groupings. However, conversations with practitioners convinced us that the approach would be too prescriptive and less useful for informing exhibit design and development in practice.
3. For those who are interested, we did check to learn the percentage of variability explained by the full set of EDGE Design Attributes for each engagement measure. Over and above the control variables (institution and crowdedness), the set of 9 explained 7% variability in girls’ Use (for a total of 71% explained), 12% variability in girls’ Return Visits (for a total of 31% explained), 9% variability in girls’ Time Spent (for a total of 31%), and 20% variability in girls’ EL scores (for a total of 29% explained).

4. Appendix A also includes the confidence intervals, for those interested in the precision of the results.

Acknowledgments
This research was made possible by the hard work, creativity, and deep thinking of the staff from all three science centers, as well as our project team, advisors, and reviewers. We are appreciative of the museum visitors who agreed to participate in this research. We are grateful for the support of the National Science Foundation. 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


Cosmos Corporation. 1998. *A report on the evaluation of the National Science Foundation's informal science education program.* National Science Foundation.


Dancstep (née Dancu), T., and V. Garcia-Luis. Forthcoming. Straight from the girls: Why you should incorporate the EDGE design attributes at exhibits.


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


Serrell, B. 2010. Paying more attention to paying attention. *Center for Advancement of Informal Science Education*.


This article is protected by copyright. All rights reserved
Table I. Sample Female-Responsive Design Strategy and Adaptation to Exhibit Design.

<table>
<thead>
<tr>
<th>Female-responsive strategy</th>
<th>Adaptation to design</th>
<th>Exhibit design attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enable Social Interaction and Collaboration</td>
<td>Provide enough space to accommodate a friend or a group</td>
<td>The exhibit is designed to allow people to experience the phenomenon or do the activity from more than one side</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The exhibit includes more than one station</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The exhibit has a large floorplan that is designed to accommodate three or more people</td>
</tr>
<tr>
<td></td>
<td>Encourage discussion</td>
<td>The label invites visitors to compare with another person</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The label recommends telling/showing others</td>
</tr>
</tbody>
</table>

Table II. Participant demographics.
### Demographic Information

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

a. Gender was equally distributed at each of the institutions. However, attendance and race/ethnicity varied by institution.

b. Ethnicity totals for each category include multi-racial/multi-ethnic participants; thus totals per category may be greater than participant totals.

### Table III. Engagement Level rubric.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Examples (may do one or more of these)</th>
</tr>
</thead>
</table>
| 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 themselves.

- 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.)

### Table IV. The final EDGE Design Attributes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The exhibit’s look-and-feel is playful, whimsical, or humorous.</strong> The central experience or aesthetic fosters a feeling of playfulness rather than a need to “be serious” or “get it right.”</td>
<td>Significant $^1$ $^2$ 5% more $(sr = .15)$</td>
<td>Significant 6% more $(sr = .23)$</td>
<td>Significant 13.5 seconds more $(sr = .15)$</td>
<td>NS 2% more $(sr = .12)$</td>
</tr>
<tr>
<td><strong>The exhibit’s look-and-feel is homey, personal, homemade, or delicate.</strong> Homey design aesthetics deal with materials such as soft fabrics, wood cabinetry, or a small, intimate scale.</td>
<td>NS 2% more $(sr = .04)$</td>
<td>Significant 5% more $(sr = .17)$</td>
<td>Significant 15.2 seconds more $(sr = .15)$</td>
<td>Significant 6% more $(sr = .30)$</td>
</tr>
<tr>
<td><strong>The exhibit has multiple stations or sides, allowing more than one person to experience the phenomenon.</strong> These exhibits give each visitor ownership of a personal space to experience the phenomenon or do the activity.</td>
<td>Significant 7% more $(sr = .22)$</td>
<td>Significant 5% more $(sr = .20)$</td>
<td>NS 7.8 seconds more $(sr = .10)$</td>
<td>NS 1% more $(sr = .06)$</td>
</tr>
<tr>
<td>The exhibit is open-ended, providing multiple outcomes, activities, or ways to interact.</td>
<td>NS</td>
<td>Significant 2% more ((sr = .06))</td>
<td>Significant 19.4 seconds more ((sr = .24))</td>
<td>Significant 5% more ((sr = .25))</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Exhibits may be open-ended in many ways, for example: the outcome is different every time or it is designed for a multitude of iterations with an assortment of variables.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The exhibit has been designed with space to accommodate three or more people. A large room, spacious floor plan, or large table surface create physical space, or elbow room at the exhibit.</td>
<td>Significant 5% more ((sr = .15))</td>
<td>Significant 4% more ((sr = .18))</td>
<td>Significant 11.2 seconds more ((sr = .14))</td>
<td>NS 1% more ((sr = .06))</td>
</tr>
<tr>
<td>The exhibit label includes at least one image of a person. An exhibit label might include an image of a person to add real-world context or help visitors use the exhibit.</td>
<td>NS 2% more ((sr = .06))</td>
<td>NS 1% more ((sr = .06))</td>
<td>Significant 9.5 seconds more ((sr = .12))</td>
<td>Significant 4% more ((sr = .26))</td>
</tr>
<tr>
<td>The exhibit is designed so visitors can watch others to preview what to do. Visitors can see the actions or reactions of another person using an exhibit.</td>
<td>Significant 5% more ((sr = .13))</td>
<td>Significant 5% more ((sr = .19))</td>
<td>NS 5.5 seconds more ((sr = .06))</td>
<td>NS 1% more ((sr = .03))</td>
</tr>
<tr>
<td>The exhibit label includes a use drawing, giving visitors an idea of how to use the exhibit. These drawings often show a person doing an action, or how to use an interactive element.</td>
<td>NS 1% more ((sr = .03))</td>
<td>NS 0% more ((sr = .01))</td>
<td>Significant 8.7 seconds more ((sr = .10))</td>
<td>Significant 8.7 seconds more ((sr = .10))</td>
</tr>
<tr>
<td>The exhibit includes at least one familiar object that most people have seen before. Includes everyday things such as kitchen items, household tools, musical instruments, or stuffed animals.</td>
<td>NS 1% more ((sr = .04))</td>
<td>Significant 3% more ((sr = .10))</td>
<td>NS 7.5 seconds more ((sr = .08))</td>
<td>NS 2% fewer ((sr = -.08))</td>
</tr>
</tbody>
</table>

1 Significant: \(p < .05\); NS: \(p > .05\)

2 Semi-partial correlations \((sr)\) provide the unique relationship between each design attribute and each engagement measure, controlling for crowdedness and institution. According to Cohen (1988, 1992), .10 is considered small but meaningful, .30 is considered moderate, and .50 is considered large.
Table V. The design attributes that were no longer significant after controlling for institution and crowdedness.

<table>
<thead>
<tr>
<th>EDGE DESIGN ATTRIBUTE</th>
<th>Girls' USE RETURN VISITS</th>
<th>Girls’ TIME SPENT</th>
<th>Girls’ HIGHEST ENGAGEMENT LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>The exhibit has bright, prominent color.</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Color is central to the phenomenon, or a highly visible part of the interactive elements, cabinetry, hardware, or label. This attribute looks at the overall salience of color at the exhibit; it does not evaluate specific colors.</td>
<td>0% more ($sr = .01$)</td>
<td>1% more ($sr = .05$)</td>
<td>.3 seconds less ($sr = -.00$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2% fewer ($sr = -.12$)</td>
</tr>
</tbody>
</table>
Exhibits in study (n = 334 exhibits)

- Excluded (n = 32) Location confounds
- Excluded (n = 33) Broken/removed exhibits Too few users
- Excluded (n = 22 exhibits) Short use is expected
- Excluded (n = 220 exhibits) Low reliability
- Use Analysis (n = 269 exhibits)
- Return Visits Analysis (n = 301 exhibits)
- Time Spent Analysis (n = 279 exhibits)
- Highest Engagement Level Analysis (n = 81 exhibits)

This article is protected by copyright. All rights reserved
CORRELATIONAL ANALYSIS

Exhibit design attribute data

- Excluded attribute if:
  - Low sample of exhibits

- Excluded attribute if:
  - Negative correlation with engagement (any inst./any measure)

  - Positively correlated at only one institution AND
    Positively correlated for only one engagement measure

  - Redundant

Girls’ engagement data

- Use
- Returns
- Time spent
- EL Analysis

REGRESSION ANALYSIS

Controlled for:

- Crowding
- Institution

EDGE Design Attributes