Visitor Studies

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uvst20

Using Embedded Assessment to Evaluate Science Center Theatre

Joe E. Heimlich, Laura Weiss, Jacob Cason, & Whitney Owens

To cite this article: Joe E. Heimlich, Laura Weiss, Jacob Cason, & Whitney Owens (2024) Using Embedded Assessment to Evaluate Science Center Theatre, Visitor Studies, 27:2, 144-159, DOI: 10.1080/10645578.2024.2343583

To link to this article: https://doi.org/10.1080/10645578.2024.2343583

VISITOR STUDIES
2024, VOL. 27, NO. 2, 144–159
https://doi.org/10.1080/10645578.2024.2343583

Routledge
Taylor & Francis Group

Using Embedded Assessment to Evaluate Science Center Theatre

Joe E. Heimlich[a], Laura Weiss[a], Jacob Cason,[b] and Whitney Owens[b]

[a]Center of Science and Industry, Columbus, Ohio, USA; [b]Cincinnati Museum Center, Ohio, USA

ABSTRACT

While some people claim museum theatre is a powerful tool that can affect visitors' attitudes and inspire learning, others continue to question if its role is education, entertainment, or some combination of the two. In light of these mixed perceptions of museum theatre, there is ongoing demand for evidence that these shows promote meaningful learning. One method for measuring outcome achievement in these nontraditional theatre settings is embedded assessment (i.e., unobtrusive instruments integrated into the experience being evaluated). This article presents a case study of using embedded assessment to measure the stability of audience understanding of three key concepts from a science show. During *Lunch in Space*, the project team developed a strategy where the audience took on the role of food scientists and completed a paper and pencil "menu" by selecting good foods to send into space. Results suggest that this is a stable strategy for measuring short-term learning from this type of program. Three key insights learned for embedded assessment with museum theatre include: 1) creating a role for the audience; 2) refining the assessment through rapid iteration; 3) and designing assessments with a universal design approach, simple, easy-to-understand instructions, limited text, and clear visuals.

ARTICLE HISTORY
Received 5 January 2023
Revised 29 February 2024
Accepted 1 March 2024

KEYWORDS
Embedded assessment; evaluation; museum theatre; science centers

Introduction

Theatre[1] in museums is not just about performance, but can also be a pedagogical tool (e.g., Carol-Ann Burke et al., 2018; Lanza et al., 2014; Lewandowska & Weziak-Bialowolska, 2020). In museum settings, theatre takes many forms; Bridal (2004) provides a long list, from performances of scripted pieces by actors in defined spaces to educational activities presented by actors. To this list, Hughes (2010) and Kiurski (2022) add an array of others such as puppetry, storytelling, monologs, and mimicry.

Regardless of how much and how different theatre engagements may look across types of institutions, there are ongoing questions about whether museum theatre's role is, in general, one of education, entertainment, or edutainment (Moustafa, 2020; Rapeepisarn et al., 2006). Even with increasing evidence that learning can and does

CONTACT Joe E. Heimlich jheimlich@cosi.org Center of Science and Industry COSI, 333 W. Broad Street, Columbus, Ohio, USA.

happen through museum theatre (e.g., Carol-Ann Burke et al., 2018; Cohn et al., 2011; Stagg & Verde, 2019), museum theatre is continually both "valued and shunned by museum directors in almost equal measure" (Jackson, 2002, p. 6). As an educational tool, it is the context that sets museum theatre apart from other forms of educational theatre practice, and challenges museums to find a balance between the artistic and the pedagogic (Hughes, 2010).

Theatre in museums can be more than just teaching content and explaining how things are or were (Evans, 2013). The use of drama or dramatic techniques in the museum context can provoke visitors' emotional or cognitive examination of the museum's discipline (Hughes, 1998) and engage learners in creatively thinking through scientific problems and challenges (Tselfes & Paroussi, 2009).

While some people claim theatre is a powerful educational tool that can affect visitors' attitudes and inspire learning, others feel little research or critical evaluation has been conducted to defend those claims (Malcolm-Davies, 2007). Some of the problems with the measurement of museum theatre include simplistic self-reporting measures, simple content knowledge gains, few attempts to segment or compare different audiences, and little research into how venue, context, or performance quality influences outcomes (Austin & Sullivan, 2019). However studies that do exist suggest there is mission value in museum theatre as shows can increase knowledge and positive attitudes (e.g., Schechter et al., 2010; Price et al., 2015) and motivate visitors in many ways (Walker, 2012).

Science shows tend to be somewhat different from museum theatre in other types of museums as they tend to incorporate more demonstration and less character or narrative (e.g. DeKorver et al., 2017). They are usually built with two goals: teaching content knowledge, and engaging the audience with science (Peleg & Baram-Tsabari, 2011). There is evidence that science shows can be successful with both goals. For example, across 10 years of evaluation studies of science theatre in one museum, evaluators note three general outcomes: 1) knowledge gain, 2) visitor perception of plays as educational and of consequent value, and 3) visitors' articulation of abstract and complex ideas (Baum & Hughes, 2011). But even in these studies, the evaluation is conducted as a step removed from the experience itself, which creates a disruption to the natural end of the show. Both of these concerns create a challenge as intercepts are difficult when exits are in all directions and when all audience members have psychologically ended their experience and are moving en-masse toward their next experiences.

This article reflects one approach to documenting the desired cognitive outcomes of a grant-funded theatre show, *Lunch in Space*, while trying to minimize the audience's sense of testing or measurement. Keeping the challenges presented above in mind, the project team sought to answer one question: can the *Lunch in Space* audience, primarily families, apply the science concepts behind how food is selected for space travel to determine if different foods are appropriate for space? The team strove to make the evaluation experience a component of the show itself and to thereby embed the evaluation as an integral part of the visitor experience during the show. This article will first be grounded in embedded assessment, then share the methods, tools, and findings from the embedded assessment for evaluating the learning of the *Lunch in Space* show. Particular attention will be paid to the process of creating this

embedded assessment including the critical friend role (e.g. Costa & Kallick, 1993) in which the evaluator intentionally pushes hard against team decisions so that all perspectives are covered, and rapid prototyping in which changes are made in real time for iterative testing.

Background: Embedded assessment

In their classic book *Unobtrusive measures: Nonreactive research in the social sciences*, Webb, et al(1966) noted the predominance of interviews and questionnaires as the sources for social science research data. In the introduction, they "lament" the over-dependence on those two sources and suggest that while some believe going beyond those might "not leave much. It does. Many innovations in research methods are to be found scattered throughout the social science literature. Their use, however, is unsystematic, their importance understated" (p.1). Since their exploration of alternative ways of obtaining data through unobtrusive measures to replace or to add depth to the more traditionally gathered data, embedded assessment has continued to be in the background of social science research.

As defined in formal education, embedded assessment activities are recurring activities that appear indistinguishable from instructional activities, but which generate both formative feedback to students and summative feedback about student performance/reports for teachers, parents, and administrators (Kennedy et. al, 2006). There are two broad approaches to embedded assessments in schooling, through authentic projects such as repeated task performance or through informal discussions or conversations with the students. In informal[2] and nonformal settings, embedded assessment activities likewise appear indistinguishable from instructional or engagement activities to the visitor but provide feedback to the educators and the institution about the efficacy of the program or experience. This is important as Allen and Peterman (2019) note, traditional educational assessment techniques used as the standard norm in informal contexts may 1) threaten the ecological validity of informal (and even many nonformal) learning experiences, 2) undermine the nature of brief, voluntary, and emergent learning experiences, and 3) not fit within traditional bounds of a "treatment" in traditional learning contexts (pp 19-20). In any context, levels of integration of embedded assessments range from direct assessment activities that may or may not be part of a coherent scenario, to completely transparent, unobtrusive sets of actions or "stealth" assessments (Shute, 2008; Shute et al., 2009).

The concept of embedded assessments as less obtrusive than traditional tests suggests that in the informal context, visitors can engage with them without interrupting their museum experience (Fu et al., 2019). Becker-Klein et al. (2016) defined embedded assessments for community science projects as "tasks that are integrated seamlessly into the learning experience itself" (p.1), using normal program activities as the opportunity for assessment. In the informal context, these assessments are most frequently, and not surprisingly, found in technologically driven experiences. For example, online simulations are effective for science learning assessment for programs that are looking at systems and technology-rich environments (Bennett et al., 2007). In an online IT program for girls, the evaluation used an evidence-centered assessment design (Koch et al., 2009) which put forth the assumption that valid inferences about progress

toward learning goals are dependent on a coherent, well-supported assessment argument about what learners know and can do. The argument is then supported by evidence, including systematic analysis of learner performance on tasks eliciting knowledge and skills (Mislevy, 1994; Mislevy et al., 1999). Another placement for embedded assessments in informal contexts is in a game-based learner response system (Obery et al., 2021) such as having the choices made within the game context capture or demonstrate intended knowledge use or behavior outcomes.

Embedded assessments are appropriate for informal contexts as they gather evidence of participants' knowledge, skills, and/or attitudes (KSAs) in unobtrusive ways, support sporadic interactions, and maintain high levels of participant engagement (Zapata-Rivera, 2012). Methods include both performance assessments where participants demonstrate knowledge and skills, and authentic assessments which require the learning tasks to reflect real-life problem-solving situations (Becker-Klein et al., 2016). Allen and Peterman (2019) discuss unobtrusive instrumentation, which includes embedded assessments in online gaming or digital experiences and also covert data gathering with activities in which visitors create content based on embedded prompts in video or audio recording that is part of the experience.

Wilson and Sloane (2000) identify four transferable principles that guided the development of the embedded assessments using the BEAR (Berkeley Evaluation and Assessment Research Center) Assessment System, a formal education tool for guiding embedded assessments. The first principle is using a development perspective of learning to look at change over time. Following, evaluators ensure a match between instruction and assessment, management and responsibility of the teacher, and then quality evidence. These principles do not transfer evenly to informal learning contexts, but their underlying understandings do transfer: the assessments must discern levels of knowledge or ability, link assessment tasks to progress/gain, be different for different outcomes, give educators a role in fidelity of their use, and generate high-quality evidence. Fu et al. (2019) help align these principles with informal and nonformal settings by recommending that when planning an embedded assessment, evaluators should pay attention both to its directness (how closely the measure approximates the outcome in practice), and its obtrusiveness (how much the measure intrudes on the visitor experience). Fu et al. overlaid two continua, whether the assessment is less or more obtrusive, and less or more direct. This results in four domains with the desired state being less obtrusive and more direct.

Overall, the literature on embedded assessments in informal and nonformal learning contexts is sparse. Embedding assessments in informal contexts is a challenge for several reasons. One is the very nature of learning outcomes requiring the visitor to demonstrate their knowledge, response to, or ability from the learning in the activity. This assumes visitors share the experience in similar ways. Without the classroom construct of written work and practice exams, embedded assessment must take on a very different appearance. Creating authentic activities that generate data is often a challenge in these contexts. Perhaps another reason for the scarcity of studies of and using embedded assessments is a function of the inability to report "rigor" in those measures as are expected from the traditional tools used in research and evaluation. While the articles found in our exploration of the literature for this project and this

article used games, creative tracking and timing approaches, and an array of interactives to increase engagement in evaluation (and research) studies in informal contexts, there were relatively few embedded assessment studies. To that end, this particular study is exploring the process of embedding a measure with the intention of ensuring consistent and logical outcomes of learning through a science museum theatre production.

The process and methods

About the show *Lunch in Space* grew out of the larger NSF-funded project Food for Thought: Igniting, Engaging, and Measuring Family STEM Learning Using a Food Lab. Food for Thought is a partnership between Kent State University, Cincinnati Museum Center (CMC), and LeSoupe that focuses on engaging families with content and activities around food science and cooking. Center for Research and Evaluation (CRE) serves as the process evaluator for the project. Like many recent projects, the grant's original plans were affected by the COVID-19 pandemic and the Food for Thought team had to reevaluate their plans, critically revise how and how many participants might be possible, and consider other ways to interact with families around food science with the least direct contact. As the pandemic moved into a return to museum visits, the idea for using stage shows at CMC as a vector for conveying messages emerged. The team's goal was to convey cognitive content through the show in a way that leads to demonstrable learning. More importantly, the team built both the show and the assessment through a collaborative process to ensure the outcomes were clearly and consistently presented in the show, and that the assessment measures if the basic food science constructs were transferred to the learner.

The iterative process

During a full project team meeting, the CMC team identified space as a topic with many potential connections to food science and strong appeal to museum audiences (necessary for attracting visitors to watch a stage show). This idea took root and, in the discussion, the idea of embedding assessment into the show to try to circumvent the challenges of traditional evaluation approaches was discussed. The evaluation team offered ideas and the rest of the team added suggestions of what could be. It was in the collaborative brainstorming that the idea for an assessment format and framing emerged. From this meeting, a process was determined for moving forward.

CMC educators wrote a script that became the initiation point for the iterative refinement of the show and embedded evaluation. *Lunch in Space* is a 15-minute science stage show that invites visitors to take on the role of NASA food scientists and help plan a menu for the International Space Station. The presenter employs props, videos, and humor to make the show visually exciting and fun for a family audience. In terms of content, messaging centers on three key themes for identifying if foods are good or bad for space: crumbs, weight, and liquids.

The first iterations of the script were focused on the show elements and exploration of what key concepts about food science and space would be most appropriate for the many younger members of the show's audience while still engaging and teaching the adults. Alongside the development of the script, the CMC team tried several different

ways of creating the embedded assessment menu. This included testing different foods to see which ones appeared to be the best indicators of the transfer of learning.

The CRE team, in their critical friend role (e.g. Costa & Kallick, 1993; Handal, 1999), read the first version of the script, and offered critical feedback and alternative perspectives, especially around the consistency of messaging, framing of the messages, and maintaining focus on the intended outcomes. The museum team guaranteed that the script would be high energy, include a lot of humor, and incorporate some of the more typical "whiz-bang" moments science theatre shows are known for. Subsequent versions of the script were performed even as the museum team continued testing different foods on the menu. Changes to the script also led to changes to the menu (discussed further below).

After several iterations and tests, the CRE team visited the CMC and observed four shows. The observation checklist included a capture of the number of times major themes were addressed and the various ways in which they were expressed (to maximize learning), how the audience was brought into their role as food scientists, and specific mentions of food science and science generally. The evaluators also conducted short intercept interviews with adults exiting the show to capture key takeaways and impressions. After each performance, the evaluation team met with the museum team. Using the observational and interview data, the evaluators talked through what they saw, where the strengths of the performance and the script lay, and potential areas for improvement. The presenter, the CMC Food Science Coordinator, and the other CMC team members then discussed and set changes. This happened for each of the four shows observed.

About the embedded assessment

Creating the embedded assessment itself—named the menu– was also a collaborative and iterative process between the CMC and CRE teams. Together, the teams strove to find foods not otherwise mentioned during the show that would be both good and poor choices for space based on the final three decision criteria of crumbs, weight, and liquid. The CMC team designed and distributed the menu as a paper/pencil measure. In total, there were three iterations of the menu before reaching what became the nearly final version. Each menu had brief written instructions and labeled, color images of around 10 food items. After trying each version across multiple shows, the team discussed the response patterns, starting with the lower-scoring items—were these the more difficult items (i.e. intentional challenge for audience members) or was there some other reason affecting responses (e.g. picture quality, audience familiarity)? The team changed out images of food items to avoid confusion, similarity to another menu item, or unclear pictures until the conceptual reason for making the food decision was clear.

The two items with the fewest correct responses across versions of the menu were freeze-dried ice cream (labeled as astronaut ice cream) and popcorn—both included as intentional challenges for audience members. The challenge with the freeze-dried ice cream was to look at the picture, rather than the name, and notice crumbs (making it a bad space food). During the first observation debrief, it was noted the presenter should not call it a "trick" for the audience, but instead explain how something well-known as 'astronaut food' is actually misleading. With popcorn, the challenge is

the difference between how it is eaten on Earth (in handfuls) versus in space (piece by piece, which avoids crumbs and makes it a good space food). Also during observation debriefs, the team recommended that the presenter talk about other foods astronauts eat in single bites (e.g., M&Ms, which were already a prop in the show to discuss packaging foods for space). Even if visitors continued to struggle with this connection—as it is a much different way than they are accustomed to eating popcorn—the post-menu debrief was another opportunity to reinforce messaging about ways astronauts avoid making crumbs.

For each version of the menu, the team critically looked at each of the items and discussed whether they "best" illuminated the decision criteria. For example, the first menu version included four items tied to liquid, but the number of correct responses was low for yogurt and mashed potatoes. With the script focus on not creating crumbs, many audience members thought these liquids were good space foods, rather than recognizing them as liquids and high-weight foods (i.e. space-unfriendly). Another food item that did not work in this context was crab legs. The team originally chose the item because cracking the shell creates crumbs, but many visitors did not make this connection. Upon reflection, the team acknowledged that visitors in Cincinnati, Ohio are likely not eating fresh crab legs regularly. In the final version of the menu, the team replaced crab legs with ballpark peanuts, a more familiar shelled food item. After discussion and identifying possibly better food item selections, changes were made to the menu to improve usability.

Another example of a change that was made to ensure responses were more accurately interpreted on the menus was simplifying the instructions and reducing the amount of text on the menu. the first two versions of the menu asked visitors to circle items as either 'space food' or 'earth food' while the third and fourth versions instead had visitors circle the pictures of the foods they thought 'would be good space food.' Other improvements toward usability included increasing the size of the text and images for consistency, readability, and visual appeal.

The intention behind the embedded assessment was to engage those viewing the show in the narrative and incorporate a task that would provide data for gauging key learning points while feeling authentic and natural in the moment. So, while visitors completed the menu at the end of the show, the presenter helps frame the task as an activity throughout the script. The presenter solicited learning responses from audience members at various points during the show (e.g., asking what would happen to a can of soda in space). The presenter emphasized the role of the audience throughout the script, especially in later iterations. At the beginning of the final version of the show, the presenter tells the audience that they will be NASA food scientists and help "plan a meal and select items for our astronauts." Framing the audience as food scientists is both inviting for visitors and helps prepare them for scientific thinking and receiving key messages. Near the end of the show, a second CMC team member distributed the menus and prompted visitors to circle good space foods. After collecting the completed menus, the presenter continued the show by reviewing each of the menu items and having the audience help explain why items were or were not good space foods. This added to making the activity feel like part of the show (as opposed to an exit survey) and continued to reinforce visitors' learning and the show's key messages.

There was a very simple observation checklist created that listed on the left the key outcome messages for the show and had one column for capturing occurrences of the message and another for capturing the ways messages were shared (e.g., the varying words/phrases used). After each show the evaluators would average their numbers and look across the examples to prepare summary comments for the CMC team.

A two-question intercept interview was constructed. Three evaluation team members would intercept as many adults as they could as visitors were exiting the theatre to ask 1) What three words would you use to describe the show you just watched? and 2) Why did you choose those words? The evaluators talked through what they heard and collectively summarized the findings in real time to include in the feedback session.

How data informed the iterative process: Findings

The first show observed had the fewest occurrences of the key messages (i.e. crumbs, weight, and liquids), with an average of 26 between the three observers. The second show, which occurred shortly after a team debrief about the first show, had the most occurrences of the key messages, with an average of 82. The third and fourth shows, which were observed on a different day than the first two, both had an average of 53 occurrences. The counts of the individual messages were not identical across the two presentations, but the overall frequency being similar to each other and in the middle of the first two scores suggests a stable message delivery. The show itself must be flexible enough to change based on audience interaction, so having a consistent narrative that was conceptual rather than verbatim was important. Team discussions of the observations also generated ways to weave the messages of crumbs, weight, and liquids together throughout the show instead of presenting them separately. For example, crumbs and liquids can cause similar problems of getting into the machinery on the Space Station, and liquids can contribute a significant amount of weight. Sharing these challenges together rather than separately helped strengthen visitor comprehension.

The comments received from the adults interviewed as they were exiting the show were intentionally brief. The evaluation team simply wanted to know what top of mind messages visitors were taking away from the show. While limited as the three researchers could only obtain data from one or two adults each as people were exiting rapidly from the theatre space, the data were useful in helping hone some of the language in the script. The primary outcome messages were usually named by at least a couple of those interviewed and other comments revealed insights that reinforced script decisions, such as not having thought of food science as having an important role in space and the show revealing a career in science the individual had never before considered. These data were immediately provided to the museum team for integration and solidification.

As this article is focusing on the process, findings were primarily used to determine if the museum team's desired level of correctness was consistently achieved, which, based on the age range and distribution of respondents, the CMC set at around 2/3 of respondents. Across the first four versions of the menu, there was an overall correct response percentage of 64.9% ($n = 492$). The percent correct was similar in the first two versions (51.8% and 54.2%, respectively), and again in the third and fourth versions (64% and 66.1%, respectively) (see Figure 1). The overall mean was slightly depressed by the first two versions, but there were fewer subjects in those tests ($n = 20$

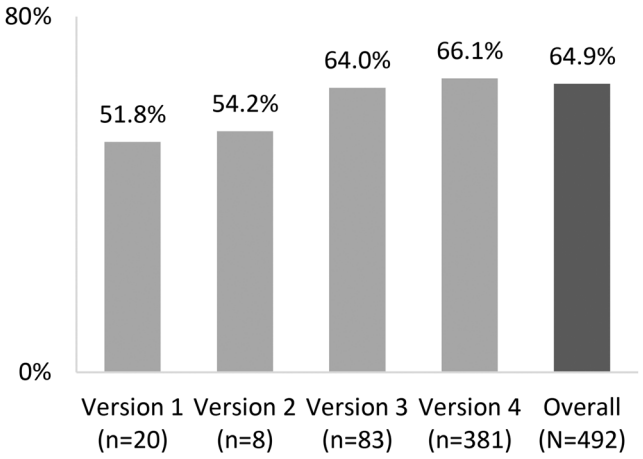

Figure 1. Overall percent correct by menu version (Heimlich & Weiss, 2023).

and $n = 8$, respectively). The seemingly lower percentage of correct responses in these versions is likely due to both the smaller number of respondents and the earlier versions being more preliminary and exploratory. The improvement in both the show and the measure likely led to the shift to the higher means of the third and fourth versions. The consistency in the first two versions and then especially in the second two versions, suggests the iterative process worked in increasing the transfer of the concepts to other foods and that the menu could provide a stable strategy for measuring short-term learning from this type of program.

This final version of the menu was used over several weeks to gather data to explore consistency ($n = 381$). For eight of the 10 food items, a majority of the participating audience members (60-70%) chose the correct responses (i.e., circling items that are good in space and not circling items that are bad in space). For six of these items, five of which were bad space foods, over 70-% of respondents answered correctly. For the other two items, both of which were good space foods, over 60% of participants answered correctly, which was within the range of the desired target for stability in the outcome score. As mentioned earlier, the two food items that continued to be challenging for visitors were freeze-dried ice cream and popcorn, but the ability to use them to highlight the reasons they are good space foods during the post-instrument review was considered a strong pedagogical approach (see Table 1).

Table 1. Correct responses by item for version 4 of the menu (Heimlich & Weiss, 2023).

Menu item	Good or bad space food	Related key message	Number correct	Percent correct
Hamburger on a bun	Bad	Crumbs	327	85.8%
Potato chips	Bad	Crumbs	327	85.8%
Fire roasted s'mores	Bad	Crumbs	326	85.6%
Dried banana chips	Good		312	81.9%
Ballpark peanuts	Bad	Crumbs	292	76.6%
Watermelon	Bad	Weight, Liquid	281	73.7%
Granola bar	Good		249	65.4%
Peanut butter tortilla	Good		234	61.4%
Popcorn	Good		109	28.6%
Freeze-dried ice cream	Bad	Crumbs	62	16.3%

When considering the overall correct percentage, it is important to acknowledge that the show takes place in a free-choice setting and that young children through adults were completing the assessment. In this museum context, visitors can arrive late, which means they may complete the menu at the end after missing some of the show's earlier messaging. During observations, the team noted that some younger children appeared to be considering which foods they like most (e.g., potato chips) more than their appropriateness for space. Given these conditions, a stable two-thirds correct response was determined to be solid, especially considering that the "challenge" items of freeze-dried ice cream and popcorn were the ones that drew the percentage correct score down the most.

These data were used as the markers to accept the menu as a useful tool for both the pedagogy of the theatre show and the ability to engage audiences in the embedded measure of the menu. Though the study has concluded, the activity remains an important component of the show. There is also intention to periodically use the data from a show or a few shows to ensure ongoing consistency in the outcomes.

Discussion

The actions required for embedding an assessment in formal education as identified by Wilson and Sloane (2000) use a developmental/iterative process, ensure a match between instruction and assessment (pedagogy and outcome), align the management and responsibility of the educator, and ensure quality evidence—are useful when creating an unobtrusive evaluation of a museum theatre program/science show. From the initial conceptualization of the show through to the final version, the close collaboration and iterative process were vital for ensuring the menu was integral to the experience *and* was measuring the intended learning outcomes. The critical friend role of the evaluation team became important in helping capture the script components that began as ad-libs and then were incorporated into the script as regular components. For example, reiterating the audience role multiple times and referring to scientific phenomena outside of food science were codified through the observation and debriefs.

Another valuable component of the process was the CRE Team's role as education and theory experts – first in the brainstorming, and then in a structural review of the script. This process aligned with CMC'S education-forward approach to the script, which reflects traditional classroom lesson construction such as having three short lessons, or in this case, three key messages focused on specific learning outcomes with selected demonstrations to illustrate them (Kerby et al., 2010). This process also aligns with the findings from Peleg and Baram-Tsabari (2011), who noted when they did this type of work that museum audiences could decode and name the explicit and cognitive messages. This further aligns with Peleg and Baran-Tsabari (2011) findings that museum theatre audiences decode explicit but not implicit measures, reinforcing the need for the repetition of concepts in different ways to ensure the message is received.

This work also must be undertaken with intention. As the Museum team was developing the concepts and the evaluation team was constructing the measure, it became obvious that an embedded assessment requires a long-term commitment to the assessment—the assessment cannot be removed without changing the show. Thus,

the realization emerged that we could use the data in an iterative way to help design the show and the narrative, and in a more summative manner to ensure the learning was present as intended. The menu also continues on, providing visitors an opportunity to test their knowledge, and the museum an opportunity for periodic examination to ensure message stability continues.

To be truly embedded, the measure needed to fit in the narrative of the show. Narrative is central to theatre in museums and is one of the reasons that museum theatre can be educational (Jackson, 2002; Peleg & Baram-Tsabari, 2011) as the narrative provides context and illustration in a way that helps younger learners understand key messages (Lanza et al., 2014). One component of the narrative is the way in which the audience experiences itself in the show/performance, an element critical in allowing a visitor to become a willing audience member or participant. All this tells us the embedded assessment must be incorporated into the narrative so it is authentically part of the learning exchange (Jackson, 2002; Peleg & Baram-Tsabari, 2011). Further, for the audience members to continue to engage, the choice as to whether, and how deeply, to participate is a series of important, necessarily explicit decisions that depend on how the experience is framed (Jackson & Kidd, 2007). In the particular case of *Lunch in Space,* this was established early by inviting the audience to become "food scientists working for NASA," which introduced a concept of science in space travel that few had previously considered (as mentioned in the exit interviews). The continued reference to both the role and the activity reinforced the centrality of the activity to the narrative itself.

Other theatrical elements important to youths' recollections of a play include props, stage effects, and characters (Peleg & Baram-Tsabari, 2011). For this production, the presenter was highly-trained, charismatic, and interacted well with all ages of audience members. The skill of this presenter was a major factor in the engagement of the audience with the narrative, and the rapidity with which changes were made to the script/performance in real-time. The museum ensured the quality of the props, including the space-lunch board created to replicate the meal boards used on the Space Station, and a NASA-branded Earth lunchbox l, the contents of which the presenter used to create visual effects by hanging the food on magnets to represent it "floating" in space. Further, the intermittent videos of astronauts on the Space Station doing things such as eating and playing with liquids and gravity helped further illustrate the key messages highlighted in the iteratively developed script.

Conclusions

For a theatre piece or a science show to lead to intended outcomes, this study's key findings demonstrate that practice must be grounded in good pedagogy and/or andragogy. For example, designers should use motivational features strategically to enhance key messages and enhance surprise and curiosity through schema-manipulation techniques. These techniques include asking for audience predictions, using multi-leveled demonstration structures, and creating optimum-sized info gaps for the audience to complete. Designers can enhance perceptions of value by placing the science in real-world contexts that are appealing to the specific audience.

As found across the literature, three factors are important contributors to the effectiveness of not only lessons, but also of shows - strategic use of demonstrations, emotional connection, and audience engagement.

For embedding assessments, the team gained three key insights from this work:

1. Creating a role for the audience as a character (e.g. NASA food scientists) works well for engaging the audience in critical thinking about the lessons they are learning. The assessment task (the menu) was an appropriate activity for the role of the character. Additionally, the role positioned the audience as decision-makers requiring them to think through the scientific problems each of the foods might present in space (e.g. Tselfes & Paroussi, 2009).
2. Refining the assessment through rapid iterations works well to discover biases embedded in different images/items and allows the assessment to examine critical thinking rather than rote recall. It is in the iterations that the demonstrations, emotional connections, and audience interactions museum theatre offers can be explored and enhanced (Austin & Sullivan, 2019).
3. This type of assessment works best when it includes simple, easy-to-understand instructions, has limited text and clear visuals, and is designed using a universal design approach to be appropriate and appealing across age groups. Because of the intention in the design, the activity used for the assessment continues to be used in every show. While not specifically for the data, if the institution chooses, the data could be gathered periodically to ensure the desired level of correct responses are consistent.

Embedded assessments take time to create, incorporate, and test. From this experience, both the museum and the project evaluators feel the time and effort were worth doing.

Notes

1. The literature is inconsistent in the spelling as theatre or theatre. As the dominant texts and the International Association of Museum Theatre use "theatre," and there is a secondary usage of theatre as the venue and theat**re** as the performing art, the authors chose to use the "re" spelling used by most English-speaking countries, except for the U.S., where the spelling is mixed.
2. We use the constructs of informal and nonformal following the conceptual frameworks of Coombs et al., 1973; Mocker & Spear, 1982; Fordham, 1993; and AUTHOR, 1993 where informal learning is determined by the individual, but the institution or educator sets their own outcomes in designing the experience. Nonformal, then is defined by the institution determining the means but the control to learn or not resides with the individual. We are also using the overarching "free-choice learning" label offered by Falk and Dierking (1998) which captures informal, nonformal, incidental, and everyday learning – all the non-school/training learning experiences.

Acknowledgment

The Authors wish to acknowledge the debt owed to the three PIs on this project, Bradley J. Morris, Ph.D. and John Dunlosky, Ph.D.of Kent State University, and Whitney Owens of the Cincinnati Museum Center. We also acknowledge the important contributions of performer extraordinaire, Jeb Brack.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was completed with support from the National Science Foundation (#190676).

ORCID

Joe E. Heimlich http://orcid.org/0000-0003-2586-5732
Laura Weiss http://orcid.org/0000-0002-3548-352X

References

Allen, S., & Peterman, K. (2019). Evaluating informal STEM education: Issues and challenges in context. *New Directions for Evaluation*, *2019*(161), 17–33. https://doi.org/10.1002/ev.20354

Austin, S. R. P., & Sullivan, M. (2019). How are we performing? Evidence for the value of science shows. *International Journal of Science Education, Part B*, *9*(1), 1–12. https://doi.org/10.1080/21548455.2018.1532620

Baum, L., & Hughes, C. (2011). Ten years of evaluating science theatre at the Museum of Science, Boston. *Curator: The Museum Journal*, *44*(4), 355–369. https://doi.org/10.1111/j.2151-6952.2001.tb01175.x

Becker-Klein, R., Peterman, K., & Stylinski, C. (2016). Embedded assessment as an essential method for understanding public engagement in citizen science. *Citizen Science: Theory and Practice*, *1*(1), 8. DOI https://doi.org/10.5334/cstp.15

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2007). *Problem-solving in technology rich environments: A report from the NAEP technology-based assessment project* NCES 2007-466, U.S. Department of Education, National Center for Educational Statistics, U.S. Government Printing Office, Washington, DC.

Bridal, T. (2004). *Exploring museum theatre*. Altamira Press.

Carol-Ann Burke, L. E., Wessels, A., & McAvella, A. (2018). Using theatre and drama to expose and expand the epistemic insights of youth regarding the nature of science. *Research in Science Education*, *48*(6), 1151–1169. https://doi.org/10.1007/s11165-018-9782-z

Cohn, S., Van Cleave, S., Onkka, A., & King, Z. (2011). *Suitcase science: The science behind people's stuff. St Paul*. Department of Evaluation and Research in Learning, Science Museum of Minnesota.

Coombs, P. H., Prosser, R. C., & Ahmed, M. (1973). *New paths to learning: For rural children and youth*. International Council for Educational Development.

Costa, A. L., & Kallick, B. (1993). Through the lens of a critical friend. *Educational Leadership*, *51*, 49–49.

DeKorver, B. K., Choi, M., & Towns, M. (2017). Exploration of a method to assess children's understandings of a phenomenon after viewing a demonstration show. *Journal of Chemical Education*, *94*(2), 149–156. https://doi.org/10.1021/acs.jchemed.6b00506

Evans, S. (2013). Personal beliefs and national stories: Theatre in museums as a tool for exploring historical memory. *Curator: The Museum Journal*, *56*(2), 189–197. https://doi.org/10.1111/cura.12019

Falk, J. H., & Dierking, L. D. (1998). *Free-choice learning: An alternative term to informal learning? Informal Learning Environments Research Newsletter*. American Educational Research Association.

Fordham, P. (1993). Informal, non-formal and formal education programmes in YMCA George Williams College ICE301. *Lifelong learning, Unit 1 Approaching lifelong learning*. London:

YMCA George Williams College. Available in the informal education archives. (https://infed.org/mobi/informal-non-formal-and-formal-education-programmes/.
Fu, A. C., Kannan, A., & Shavelson, R. J. (2019). Direct and unobtrusive measures of informal STEM education outcomes. *New Directions for Evaluation*, *2019*(161), 35–57. https://doi.org/10.1002/ev.20348
Handal, G. (1999). Consultation using critical friends. *New Directions for Teaching and Learning*, *1999*(79), 59–70. https://doi.org/10.1002/tl.7907
Heimlich, J. E. (1993). *Nonformal environmental education: Toward a working definition. The Environmental Outlook*. ERIC/CSMEE Informational Bulletin.
Heimlich, J. E., & Weiss, L. (2023). *Lunch in space: Process and evaluation of a science stage show*. COSI Center for Research and Evaluation.
Hughes, C. (1998). *Museum Theatre: Communicating with Visitors Through Drama*. Heinemann.
Hughes, C. (2010). Theatre performance in museums: Art and pedagogy. *Youth Theatre Journal*, *24*(1), 34–42. https://doi.org/10.1080/08929091003732948
Jackson, T. (2002). Between evaluation and research: Theatre as an educational tool in museum theatre. *Stage of the Art*, *15*(Fall), 6–9.
Jackson, A., & Kidd, J. (2007). Museum theatre: Cultivating audience engagement - a case study. Paper presented at IDEA 6th World Congress, Hong Kong July 21.
Kennedy, C. A., Brown, N. J. S., Draney, K., & Wilson, M. (2006). Using progress variables and embedded assessment to improve teaching and learning. Paper presented at the annual meeting of the *American Education Research Association, San Francisco, California*.
Kerby, H. W., Cantor, J., Weiland, M., Babiarz, C., & Kerby, A. W. (2010). Fusion science theatre presents *The Amazing Chemical Circus:* A new model of outreach that uses theatre to engage children in learning. *Journal of Chemical Education*, *87*(10), 1024–1030. https://doi.org/10.1021/ed100143j
Kiurski, D. (2022). Museum theatre as an educational tool: Examples in Servian museums. *Revija za Elementarno Izobraževanje*, *15*(Spec. Iss), 77–91. https://doi.org/10.18690/rei.15.Spec.Iss.77-91.2022
Koch, M., Penuel, W. R., Gorges, T.,). SHHH… Don't use the 'A' word: Embedded assessments of youth learning in informal learning environments. Haertel, G. (2009). *Proceedings of the NARST Annual Conference*. Garden Grove.
Lanza, T., Crescimbene, M., La Longa, F., & D'Addezio, G. (2014). Bringing earth into the scene of a primary school: A science theatre experience. *Science Communication*, *36*(1), 131–139. https://doi.org/10.1177/1075547012473841
Lewandowska, K., & Węziak-Białowolska, D. (2020). The impact of theatre on empathy and self-esteem: A meta-analysis. *Creativity Research Journal*, *32*(3), 237–245. https://doi.org/10.1080/10400419.2020.1821553
Malcolm-Davies, J. (2007). Borrowed robes: The educational value of costumed interpretation at historic sites. *International Journal of Heritage Studies*, *10*(3), 277–293. https://doi.org/10.1080/1352725042000234451
Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*(4), 439–483. https://doi.org/10.1007/BF02294388
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design: A brief overview*. Educational Testing Service.
Mocker, D. W., & Spear, G. E. (1982). *Lifelong learning: Formal, nonformal, informal, and self-directed*. Information Series No. 241. ERIC Clearinghouse for Adult, Career, and Vocational Education. https://eric.ed.gov/?id=ED220723
Moustafa, R. H. E. (2020). The role of edutainment in museums, Learn through play. *International Journal of Multidisciplinary Studies in Heritage Research*, *3*(1), 15–19. https://doi.org/10.21608/ijmshr.2020.180080
Obery, A., Lux, N., Cornish, J., Grimberg, B. I., & Hartshorn, A. (2021). Competitive games as formative assessment in informal science learning: Improvement or hindrance? *TechTrends*, *65*(4), 454–463. https://doi.org/10.1007/s11528-021-00619-3

Peleg, R., & Baram-Tsabari, A. (2011). Atom surprise: Using theatre in primary science education. *Journal of Science Education and Technology*, *20*(5), 508–524. https://doi.org/10.1007/s10956-011-9299-y

Pelig, R., & Baran-Tsabari, A. (2015). Understanding producers' intentions and viewers' learning outcomes in a science museum theatre play on evolution. *Research in Science Education*, *46*, 715–741. https://doi.org/10.1007/s11165-015-9477-7/

Price, C. A., Gean, K., & Barnes, H. (2015). The effect of live interpretation with theatre on attitudes and learning of children in the mythbusters exhibit. *Journal of Museum Education, 40*(2), 195–206.

Rapeepisarn, K., Wong, K. W., Fung, C. C., & Depickere, A. (2006). Similarities and differences between" learn through play" and" edutainment".

Schechter, R., Priedeman, M., & Goodman, I. (2010). *The amazing nano brothers juggling show outcomes evaluation*. Goodman Research Group, Inc.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects* (pp295–321). PA. Routledge/LEA.

Stagg, B. C., & Verde, M. F. (2019). Story of a seed: Educational theatre improves students' comprehension of plant reproduction and attitudes to plans in primary science education. *Research in Science & Technological Education*, *37*(1), 15–35. https://doi.org/10.1080/02635143.2018.1455655

Tselfes, V., & Paroussi, A. (2009). Science and theatre education: A cross-disciplinary approach of scientific ideas addressed to student teachers of early childhood education. *Science & Education*, *18*(9), 1115–1134. https://doi.org/10.1007/s11191-008-9158-2

Walker, G. J. (2012). [*Motivational features of science shows*]. Unpublished [Ph.D. Dissertation. Canberra, The Australian National University].

Webb, E. J., Campbell, D., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally & Company.

Wilson, M., & Sloane, K. (2000). From Principles to practice: An embedded assessment System. *Applied Measurement in Education*, *13*(2), 181–208. https://doi.org/10.1207/S15324818AME1302_4

Zapata-Rivera, D. (2012). Embedded assessment of informal and afterschool science learning. Short White Paper, The National Academies of Sciences, Engineering, and Medicine. SOW 06-01-12 dbasse_072564_1_-libre.pdf (d1wqtxts1xzle7.cloudfront.net)

About the authors

Joe E. Heimlich, Ph.D. serves as Sr. Director of Research for COSI, and is a researcher in COSI's Center for Research and Evaluation. He is also an Academy Professor Emeritus with The Ohio State University where he served as Leader, Environmental Science for OSU Extension, and taught in the School of Natural Resources, the Environmental Science Graduate Program, and the College of Education. His work focuses on lifelong learning about and in the environment, with interests in the integration of social role, context, and conditions of the visit as a part of an individual's learningscape. His work is primarily in informal and nonformal lifespan learning. Among other national and international service work, Joe's academic and practitioner work intersect through evaluative research, upon which he built his academic career. This work involves testing theory within daily practice through utilization-focused evaluation. Dr. Heimlich (1993) has served as President of NAAEE and the Visitor Studies Association and was on the original Scientific Committee of the World Environmental Education Congress.

Laura Weiss is a Researcher in COSI's Center for Research and Evaluation. Prior, she was the evaluation specialist at the Detroit Zoological Society, an audience Research Associate at the

Philadelphia Museum of Art, The Allen Whitehill Clowes Museum Fellow at the Indianapolis Zoo, She holds a Bachelor of Arts History from the University of Illinois, Springfield where she was in the Capital Scholars Honors Program, and a Master of Arts, Museum Studies from Indiana University-Perdue University Indianapolis.

Jacob Cason is the Food Science Coordinator at the Cincinnati Museum Center. Prior, he worked in Visitor Services at the Mississippi Museum of Natural Science. He holds a Bachelor of Science from Millsaps College in Neuroscience where he spent several years doing research and bio-ethical training. Cason's training focused primarily on clinical research with experience in social issues related to inequality. After this Cason moved to Cincinnati where he uses his neuroscience background to inform experiment design and how to discuss nutrition.

Whitney Owens is Chief Learning Officer for CMC, where she leads education, community engagement, research, collections, and exhibit content development for the Museum of Natural History & Science, Cincinnati History Museum, and Children's Museum. Owens has spent 25 years in informal science education, creating dynamic and thoughtful experiences to engage and inspire audiences of all ages. Owens graduated from Northwestern University and holds a Masters in Cultural Management from the City University of London. Past work includes leadership roles at The Field Museum in Chicago and Great Lakes Science Center in Cleveland. Owens is a Fellow with the Informal Learning Leadership Collaborative, the IDEAL Center (Inclusion, Diversity, Equity, and Access through Leadership), and Leadership Cincinnati. She serves on the boards of Cincinnati Shakespeare Company, the Women's Fund of the Greater Cincinnati Foundation, and the Greater Cincinnati STEM Collaborative.