Chapter 13
The Challenges of 'Measuring Long-Term Impacts of a Science Center on Its Community': A Methodological Review

Eric Jensen and Thomas Lister

In recent years, there have been increasing demands on informal science learning institutions to demonstrate their impacts beyond the immediate aftermath of a visit. Such research is rarely conducted because of its logistical and methodological complexity. A report commissioned by the UK government to assess whether science centres should continue to receive government support reached the following conclusion:

> We have not been able to assess whether science centres are good value for money relative to other comparator programmes. This is because there is insufficient evidence on the long term outcomes of science centres or comparator programmes (Frontier Economics, p. 2).[1]

This conclusion helped to increase the salience of long-term impact evaluation for science centers in particular, and informal science education in general.

The study by Falk and Needham (2011) entitled '*Measuring the Impact of a Science Center on its Community*' represents an ambitious effort to solve the considerable logistical, methodological and theoretical challenges inherent in long-term impact measurement of this kind. Since its publication, it has been held up as a model for informal science education impact evaluation, and widely cited for its conclusion that science centers are effective at achieving long-term impact.

Sections of this article are reprinted with permission from Wiley & Sons. Reference for the original article is as follows:.

Jensen, E., & Lister, T. (2016). Evaluating indicator-based methods of 'measuring long-term impacts of a science center on its community (comment). *Journal of Research in Science Teaching,53*(1), 60–64.

A rejoinder for this chapter follows in Chap. 14.

[1] http://sciencecentres.org.uk/govreport/docs/impact_of_science_centres.pdf.

E. Jensen (✉)
University of Warwick, Coventry, UK
e-mail: E.Jensen@warwick.ac.uk

T. Lister
University of Cambridge, Cambridge, UK

P.G. Patrick (ed.), *Preparing Informal Science Educators*,
DOI 10.1007/978-3-319-50398-1_13

It has also been touted as a best practice model for measuring informal science learning institutions' long-term impact. Subsequent studies, including a recent international impact evaluation of science center impacts, have used a similar model. As informal science educators are increasingly called upon justify the long-term impacts of their practice, it is essential to understand the current evidence and methods of conducting such evaluation. This chapter critically reviews Falk and Needham's study in detail (cf. Jensen & Lister 2016; Falk & Needham 2016), using its methodological and theoretical limitations to illustrate the issues that continue to face those attempting the difficult yet important task of evaluating the informal science education impacts.

Falk and Needham draw upon St. John and Perry's (1993) notion of an educational infrastructure to highlight the complex and multi-dimensional nature of science learning. They highlight a wide array of institutions and services that contribute to public learning and understanding of science, including formal schooling, libraries, museums, nature and science centers, aquariums and zoos, botanical gardens and arboretums, television programs, film and video, newspapers, radio, books and magazines, the Internet, community and health organizations, environmental organizations and conversations with friends and family. These institutions, services and discussions are viewed as comprising a science-learning infrastructure.

Falk and Needham's study sets out to examine the impact of one component of this science-learning infrastructure: the California Science Center in Los Angeles. Previously known as the California Museum of Science and Industry, the center was redesigned in 1993 with the expectation of a marked increase in its impact on the local public's science-related understanding, interests and behavior. The revamped Center (re)opened in 1998.

Falk and Needham's long-term impact study orbits around a growing body of research on the educational value of informal science learning institutions. For decades, these institutions have made claims about their impacts on public learning and understanding of science. However, the availability of robust impact studies supporting these assertions is limited (e.g. Jensen, 2014a). '*Measuring the Impact of a Science Center on its Community*' aims to provide a great leap forward addressing this research gap.

Falk and Needham outline two methodological approaches that they contend can be used to monitor the influence a science center has on its public's understanding of science: "inside-out" and "outside-in". 'The *inside-out* approach was designed to identify visitors to the institution and assess the short- and long-term effects that various projects, activities and exhibitions had on these visitors' (Falk & Needham, 2011, p. 2). Essentially, the "inside-out" approach entails measuring the impact of an institution through visitors who have attended and participated in its activities. This is the standard approach used in educational impact evaluations (cf. Wagoner & Jensen, 2014). In contrast, an "outside-in" approach is defined as collecting data on a population scale to examine the prevalence, incidence and outcomes of visits to a particular institution amongst different demographic categories. 'The *outside-in* approach was designed to investigate through face-to-face interviews and large-scale random telephone surveys the science understanding, awareness, and attitudes of

individuals within the broader community to determine any impact the Science Center was having on these individuals' (Falk & Needham, 2011, p. 2). The outside-in approach uses correlation analysis to ascertain differences in outcomes between visitors and non-visitors, which are then attributed to the institution. Research supporting claims that science centers and other science-related institutions are significant contributors to public understanding of science have previously employed an "inside-out" approach (e.g. Falk & Storksdieck, 2005; Falk & Gillespie, 2009; Jensen, 2014b). The study by Falk and Needham that is the focus of the present article is unique in seeking to demonstrate the alternative "outside-in" approach, and in doing so, illustrate the newly developed Science Center was having a large-scale impact on the science literacy of Los Angeles residents. The present article is therefore designed to critically assess whether this is a good model for informal science learning researchers to adopt.

The two research questions posed by Falk and Needham (2011) were:

1. Who in L.A. has visited the California Science Center and what factors best describe those who have and those who have not visited?
2. Does visiting the California Science Center impact public science understanding, attitudes, and behaviors, and if so, in what ways?

Falk and Needham (2011, p. 2) identify two major challenges that limit the validity and reliability of any approach to measuring a science centers' impact. The first challenge relates to the nature of learning per se. Science learning is cumulative, developing through a variety experiences (one of which is formal schooling) at different times during an individual's life-course (Miller, 2001, 2004; National Science Board, 2006). Falk and Needham use an individual's understanding of the physics of flight to illustrate this point. '[Ones understanding of flight] might represent the cumulative experiences of completing a classroom assignment on Bernouli's principle, reading a book on the Wright brothers, visiting a Science Center exhibit on lift and drag, and watching a television program on birds'. They rightly point out that 'no one source is sufficient to create understanding, nor one single institution solely responsible' (Falk & Needham, 2011, p. 2) in such cases. The cumulative nature of learning makes assessing the impact of a single experience (such as attending a science center) on an individual's overall understanding of science a major challenge that Falk and Needham claim to have to overcome with this study.

The second major challenge that faced Falk and Needham was to disentangle an individual institution's impact, when a wide variety of institutions make up the education infrastructure (St. John & Perry, 1993). People encounter multiple components of this infrastructure throughout their life-course, from attending secondary school to engaging with a science organization, to watching a documentary on television or visiting a museum. This complex web of institutions and services is said to provide the conditions and capacities to support science learning. Falk and Needham (2011) suggest that the collection of cross-sectional data across multiple years would overcome this challenge, allowing researchers to ascribe change over time in the public's science understanding and interest to this single institution.

Given the importance of these two longstanding research challenges that have frustrated past attempts to evaluate informal science learning impact, the present article focuses critically on examining Falk and Needham's (2011) proposed methodological solutions. This article proceeds by summarizing and critically assessing each step in the research process, from sampling to conclusions. We start by addressing the samples and their representativeness. We then evaluate the research design, including the innovations proposed as solutions to the challenges of long-term impact evaluation. We assess the details of the survey questions used to measure impact, and finally, discuss alternative approaches to evaluating long-term informal learning impacts of institutions such as science centers.

Evaluating Sampling and Representativeness

Falk and Needham's study includes survey data collected in 2000, relatively soon after the re-opening of the California Science Center, and again almost a decade later (2009). Taking these snapshots of the L.A. public's understanding and interest in science was intended to allow Falk and Needham to attribute observed changes at the population level 'to the presence of this new piece of infrastructure' (Falk & Needham, 2011, p. 3). A major strength of this study is the inclusion of non-visitors, who are so often missing from the landscape of research on informal learning institutions (Dawson & Jensen, 2011; Hood, 1995; Jensen, Dawson, & Falk, 2011). However, there are a number of major unacknowledged limitations that undermine the study's claim to have captured a representative sample of the L.A. public. Some of the study's limitations are actually revealed by Falk and Needham's (2013) paper '*Factors Contributing to Adult Knowledge of Science and Technology*', which focused on the 2009 survey data comprising the second data collection point for Falk and Needham (2011).

Sampling Description

The first sample included $n = 832$ individuals aged 18 and over. A slightly larger sample was contacted in 2009, with $n = 1{,}008$ respondents completing what was described as a comparable survey instrument. As the sample targeted the population as a whole, each sample was comprised of both visitors and non-visitors to the Science Center. Respondents were said to be drawn from five racially, ethnically, and socio-economically different communities within the Los Angeles metropolitan area: 'These communities were selected to be generally representative of the diversity of greater L.A. residents' (Falk & Needham, 2011, p. 4). The communities selected included Canoga Park, El Monte, Santa Monica, Torrance, and South Central.

Interviews were primarily conducted in English, with 14% carried out in Spanish in 2000, and 8% in 2009. Race and ethnicity were claimed to be relatively

comparable across both samples. White/Caucasian residents represented 46% of respondents in both 2000 and 2009, whilst African-American respondents represented 13% of the 2000 sample and 16% of the 2009 sample. Respondents who indicated their ethnicity as Latino/Hispanic represented 29% and 25% respectively, with a further 7% and 8% claiming to be Asian-American. Those of 'other' ethnicities represented 5% in both samples. Some of the 2009 respondents were contacted and interviewed using cellular phones, although cellular interviews made up less than 10% of the 2009 sample. Respondents' mean age of 43 was identical across both datasets. The gender distribution of respondents was skewed towards women, who made up 59% of those interviewed in 2000 and 56% in 2009. The later sample included a higher percentage of respondents who reported a household income of over $50,000/year, increasing from 44% in 2000 to 62% in 2009. This was also true for the percentage of respondents in the process of obtaining a college degree, which rose from 22% in 2000 to 28% in 2009. The number of respondents with graduate degrees also increased from 16% to 19% in 2009. While the demographics of the two samples were not totally comparable, Falk and Needham state that both samples were weighted with U.S. Census data, although they do not reveal which Census data was consulted (e.g. 2000 or 2010).

Sampling Limitations

The core claim developed by Falk and Needham (2011) is that the population of Los Angeles underwent an increase in scientific knowledge from 2000 to 2009 that can be attributed to the California Science Center's impact: 'findings from this research provide strong evidence that the California Science Center directly and significantly contributes to science learning, interests and behaviors of a large subset of the L.A. community' (Falk & Needham, 2011, p. 11). For this claim to be upheld the 2000 and 2009 samples must be equivalent; otherwise, observed changes over this 9-year timeframe may instead be attributed to other (non-impact) factors, such as increases in income and education levels between the two samples, factors clearly unrelated to the California Science Center. In this section we will raise questions about (1) whether Falk and Needham's two samples are comparable to one another and, (2) whether they are sufficiently representative samples to support population-level generalizations. We begin by questioning the equivalence of the 2000 and 2009 samples, using a detailed consideration of three variables: Ethnicity, income and educational attainment.

Upon closer examination of U.S. Census data, the claim that Falk and Needham's (2011, p. 4) samples were 'representative of the diversity of greater L.A. residents' is erroneous. One demographic category in particular was significantly underrepresented in both 2000 and 2009, making the samples unrepresentative whilst introducing a high risk of sampling bias. Hispanic/Latino residents represented 29% of respondents interviewed in 2000; nearly a decade later, this figure *decreased* to 24%. This represents a significant underrepresentation of Hispanic/Latino residents living

in L.A. during the study's timeframe. According to the United States Census Bureau, in 2000 Hispanic/Latino residents represented 46.5% of the total Los Angeles population, in 2010 this figure *increased* by 2 per cent to 48.5% of the population (U.S. Census Bureau, 2000, 2010). If Falk and Needham had used a probability sample that proportionally represented the target population, then Hispanic/Latino residents would not be so heavily underrepresented in both samples. For other researchers considering employing a population sampling approach, we recommend using a random sampling procedure stratified by key variables such as ethnicity, education and income to ensure a more representative sample.

The overrepresentation of higher-earning respondents is another indicator of sampling bias that casts doubt on the representativeness of Falk and Needham's (2011) samples. In 2000, the percentage of respondents earning more than $50,000 annually was 44%, in 2009 this figure increased substantially to 62%. According to U.S. Census data, these figures do not represent the true number of Los Angeles residents earning more than $50,000 a year. According to Census data (U.S. Census Bureau, 2000), in 2000 this figure was actually 38% of L.A. residents, and based on a 5-year estimate between 2008–2012, 50% of residents reported earnings over $50,000 (American Community Survey, U.S. Census Bureau, 2008–2012). Falk and Needham (2011, p. 4) state that 'the weighted samples were comparable […] to each other, with the exception that the 2009 sample included slightly higher percentages of respondents with higher incomes'. However, we regard an 18% increase in the number of higher-earning respondents sampled in 2009 as more than 'slight'. It means the sample skews towards higher-earning respondents, with those earning more than $50,000 a year overrepresented by 12% in the 2009 sample (American Community Survey, U.S. Census Bureau, 2008–2012). This skewness towards higher-earning respondents is a major threat to the validity of the authors' claims because, as is reported in their 2013 paper, higher income is one of the strongest predictors of self-reported understanding and interest in science. Indeed, it was later reported that respondents who 'had an annual income over US $50,000 were more likely to consider themselves as knowing a moderate amount or great deal about science and technology rather than little or nothing about these fields' (Falk & Needham, 2013, p. 441). Specifically, they found that higher-earning individuals were 1.72 times more likely than lower-earning individuals to report knowing a moderate or great deal about science and technology. Taking this into consideration, the oversampling of higher-earning respondents not only undermines the representativeness of the samples, but also brings into question the assertion that the California Science Center had a positive impact on respondents' self-reported understanding and interest in science. That is, higher income could have been a more significant factor than Science Center attendance in accounting for more positive attitudes towards science in 2009 (although statistics on the relative contribution of income were not presented in the article).

Another predictor variable that could explain part of the aggregate increase in respondents' self-reported understanding and interest in science from 2000 to 2009 is educational attainment. Respondents sampled during the second wave of data collection had obtained a higher level of education than those surveyed in 2000.

The number of respondents who had obtained a college degree increased from 22% in 2000 to 28% in 2009. Falk and Needham's 2009 sample also saw an increase from 16% to 19% in the proportion of respondents with a graduate degree. Given U.S. higher education generally requires science courses as part of the 'general education requirement' regardless of major, such educational attainment could be a confounding variable in Falk and Needham's attribution of long-term impact to the California Science Center. Unsurprisingly, Falk and Needham's (2013, p. 438) bivariate analysis of the relationship between formal schooling and self-reported knowledge about science and technology 'showed that those with a higher level of education felt they were significantly more knowledgeable about these fields'. In addition to gaining enhanced exposure to science learning, those who have obtained a higher level of education may also take a greater interest in science-related news items and programmes, increasing their exposure to sources of science learning in various ways that contribute to the self-perception of understanding science.

Shifting to the bigger picture, there is reason to question whether Falk and Needham's sampling approach yielded probability samples that would support their generalizations about the 'L.A. public', 'L.A. adults' and 'those in the L.A. area'. We can begin here with the question that determines whether a probability sample has been achieved: Did all adult residents of Los Angeles have an equal probability of being selected for participation? Clearly they did not, as everyone living outside the five selected communities within Los Angeles had a 0% chance of selection. The severity of the sampling bias incurred by only sampling these five communities could only be estimated by knowing precisely how closely these communities' characteristics align with the general L.A. population. However, Falk and Needham (2011) do not provide these details. What makes this five-communities sampling method more problematic is that a nine-year period in a diverse city such as Los Angeles is likely to see significant population turnover at the level of individual communities. This makes it more likely that the 2000 and 2009 samples are incomparable. Moreover, as can be seen from the examples discussed above, there is ample basis for skepticism about the representativeness of each of these two samples as well.

Evaluating the Research Design

The majority of existing research literature evaluating informal learning institutions relies heavily on post-visit self-reports as the main mechanism for measuring impact. However, self-reports are a particularly fraught method for this kind of impact measurement, as even the most reflexive of individuals would have great difficulty accurately self-assessing the impact of encountering one component of the science-learning infrastructure, as well as identifying a specific source from which their knowledge or interest in science was derived. Many of the cognitive biases affecting such autobiographical memory are well established in the methodological literature (e.g. Tourangeau et al., 2000).

In order to measure the impact of visiting the Science Center respondents were asked to indicate their level of agreement with 4 'impact' statements using a Likert scale from 1 (strongly disagree) to 5 (strongly agree):

- I learned one or more things that I never knew before
- My understanding of things I already knew was strengthened or extended
- I came away with a stronger interest in some areas of science or technology
- It changed my attitudes or behaviors to be more positive toward science and technology

Measurements were taken across both data sets, and results indicated that almost every adult who visited the Science Center agreed that a visit resulted in an increased understanding of science and technology. Respondents in the 2009 sample were 'significantly more likely to agree that as a result of visiting the Science Center, they learned one or more things that they did not know before, their understanding of things that they already knew was strengthened, and their attitudes or behaviors were more positive towards science or technology' (Falk & Needham, 2011, p. 7). Respondents' level of agreement between 2000 and 2009 increased for three of the four impact statements, although the mean level of agreement for the statement, "I came away with a stronger interest in some areas of science or technology" decreased from 3.97 to 3.55. In response to an additional 14 items that were added to the 2009 survey, an overwhelming majority (95%) of respondents agreed with the statement: "my understanding of science or technology was strengthened or extended by my visit to the California Science Center". Other statements recording impact included: "my curiosity about science and technology was increased by visits to the California Science Center" (85% of respondents agreed), and, "I learnt at least one thing about science or technology that I never knew before" (94% agreed). Falk and Needham (2011, p. 10) describe how 'most of these respondents also reported increases in other dimensions of science and technology learning, including increases in the affective dimension of curiosity, interest, and appreciation'.

The Limitations of Self-reporting Impacts

Many perfectly good survey questions involve requests for respondents to self-report information. These questions ask respondents to access their memories, feelings or thoughts, edit that information internally, and then select a response option from the survey form. Some self-report questions are perfectly reasonable, for example: 'How satisfied are you with your visit to the science center?'. This self-report question is appropriate because a respondent could be expected to have existing views to report. Poor quality self-report questions, however, ask respondents to conduct self-assessments of their own characteristics and capabilities that they could not reasonably be expected to judge accurately. Self-report questions can also be problematic when they require respondents to be self-aware and undergo a complicated internal editing process. For example, questions asking, 'Did you learn

anything during your visit to the science center today?' ('Yes' or 'No') would require visitors to (1) call up memories of the entire visit, (2) identify moments from that visit in which new information was acquired and, (3) identify that acquired information as 'learning'. This may be an unrealistic expectation of the respondent, inflating the likelihood of errors (deviation between what actually happened and its representation in survey data). Using self-reports as a proxy for measuring learning outcomes also suffers from the followings flaws:

- *Low in validity*. While this question does measure something (e.g. self-confidence relating to science and technology topics), it does not measure its intended concept of actual science and technology knowledge.
- *Low in reliability*. Science and technology are multi-faceted domains, encompassing thousands of different sub-domains, fields of practice and particular technologies. When one person thinks of "science", they might be thinking of human cloning or neuroscience. Another person's mental representation of "science" might focus on earthquake detection or climate change (or simply a man with white hair in a lab coat!). Given this range of representations, the most high profile, recently mentioned or personally familiar aspect(s) of science and technology would likely become the basis for a respondents answer. This means that respondents are each essentially answering different questions, depending on which aspects of science and technology are most prominent for them.
- *Bias risks overestimating knowledge*. Social desirability (and ego) may drive some respondents to overestimate their knowledge.
- *Bias risks underestimating knowledge*. Some respondents may not recognize their knowledge as "knowing something" (e.g. it may just be taken-for-granted as "the way it is") or being about science and technology. For example, they might have in-depth knowledge about why and how their heating unit works at home, but not recognize such knowledge as relating to science and technology.

Beyond the general limitations in the structure of the impact measurement approach employed by Falk and Needham (2011), the specific survey questions used also deserves close scrutiny.

Respondents were asked to indicate their level of agreement with items such as, 'my understanding of science or technology was strengthened or extended by my visit to the California Science Center' and 'my curiosity about science or technology was increased by visits to the California Science Center'. The former question is quadruple-barreled as it forces four different pathways into one question: (1) science understanding strengthened, (2) science understanding extended, (3) technology understanding strengthened, and (4), technology understanding extended. The second question is double-barreled with the inclusion of both 'science' and 'technology', but it also introduces further ambiguity by referring to multiple 'visits' (plural). For example, it is unclear how someone should answer if they felt that on one visit their curiosity increased, but on others it did not (or even declined). The other two outcome statements are also problematic. Among other limitations, the statement, 'I learned at least one thing about science or technology

that I never knew before', leaves open the risk that a respondent 'learned' something incorrect that is being counted here (e.g. 'I learned that global carbon emissions are making the planet's climate more stable'). Similarly, we have no way of knowing if agreement with the following statement is actually positive, as it is too ambiguous: 'after visiting the California Science Center, I found myself thinking about some aspect of science or technology'. For example, if people found themselves thinking about the incomprehensibility of some aspect of physics due to baffling explanations they encountered in the Science Center, they could accurately agree with the above outcome statement. The self-report survey questions used by Falk and Needham to measure learning outcome provide a clear example of the misuse and poor practice of survey of survey design.

Acquiescence Bias

Beyond such straightforward question design flaws, taken as a whole the outcome statements used in this study also introduce the risk of *acquiescence bias*. It has long been established in survey methodology that when respondents are given cues such as the 'implied direction of the question' and previous questions, responses can be 'biased by acquiescence (the tendency to agree)' (Tourangeau et al., 2000, p. 5; Cannel et al., 1981). That is, when there is a whole series of positive statements about an object, it signals to respondents that the researchers are expecting or hoping that they will agree with those statements. Indeed, prior methodological research has shown a clear tendency for people to agree with Likert scale statements. Furthermore, respondents who perceive researchers as being of a higher social status will, out of social convention or courtesy, endorse any assertion made in question, regardless of its content (Krosnick, 1999). This bias can be avoided by reverse coding half of the questions. For example, 'I enjoyed my experience visiting the Science Center' could be reversed to, 'I found my visit to the Science Center unpleasant'.

Failure to follow this basic principle of survey design makes research using similar Likert scale items susceptible to acquiescence bias. Yet, there is further reason to be skeptical regarding this particular study's findings. Methodological research has found that status differential in the form of lower social status a common cause of survey acquiescence: 'The lower the status of the respondent, as measured by the occupation of the head of the household, the greater the frequency of acquiescence' (Lenski & Leggett, 1960, p. 465). Indeed, Falk and Needham's (2011, p. 7) own analysis revealed that 'lower income respondents were […] significantly more likely to agree with most statements, especially about the Science Center providing new ideas or techniques, changing attitudes about science or technology'. However, this was not recognized as a possible sign of acquiescence bias by Falk and Needham. Further methodological research found respondents completing surveys via telephone were also more likely to exhibit acquiescence than respondents participating in face-to-face interviews (Calsyn, 1992), casting further doubt on any inferences drawn from these results.

Reporting Impact on Behalf of Another

A related, but further fraught practice is to ask respondents to report on another person's knowledge, feelings or values. This is a common problem in research that asks parents or teachers to report on the experiences, attitudes or knowledge of their children or pupils, rather than collecting data directly from the children themselves.

Falk and Needham (2011) sought to measure the Science Center's impacts on children by asking parents to assess and report on cognitive and affective outcomes. Parents were asked to indicate whether their child had obtained an increased understanding of science and technology after visiting the Science Center. They were also asked to report on their children's development of appreciation for science and whether the Science Center experience had enhanced their children's chances of future success. Parents generally agreed with the positive statements about the impact the Science Center had on their children, with 87% reporting that the visit had increased their children's understanding of science and technology. 45% believed the visit had increased their child's understanding "a lot". Apart from the obvious ambiguity and unreliability in expecting different parents to judge what counts as "a lot" of learning, it appears parents were asked to provide a single assessment of whether learning had occurred for all their children: what if one child learned "a lot", another "a little" and a third "nothing"? Are parents really likely to be making a considered judgment here? Asking parents to provide an off-the-cuff assessment of their child's learning is even more prone to error than expecting them to accurately judge their own learning outcomes. Assessing another individuals' comprehension and storage of new knowledge encountered at the Science Center is an unrealistic expectation of respondents, and an unreliable method of evaluating children's understanding of science and technology.

Falk and Needham reported that 80% of parents agreed that there was an increase in their child's appreciation toward science and technology due to visiting the Science Center. However, Falk and Needham provide no evidence of how they operationalized this outcome or ensured that 'appreciation' was interpreted in a similar way by the various parent respondents. Even if parents did have a shared understanding of this outcome, appreciation is an internal psychological phenomenon that, in this instance, only the child could be expected to self-report with any degree of accuracy. The same goes for the other child-oriented questions reporting on increased 'curiosity', 'inspiration' and 'interest'.

Even more unrealistically, respondents were asked to report whether they believed a visit to the Science Center enhanced their child's chances of future success in life (79% agreed that it did). Clearly it is impossible for parents to know whether a visit to a Science Center has increased a child's life chances, and it is poor survey research practice to ask questions about content that respondents could not reasonably be expected to know. Therefore, this instance of reporting on behalf of another's future life chances is much more likely to represent survey response biases of the kind discussed previously than any meaningful Science Center visit impact. In sum, asking parents to assess the cognitive and affective outcomes of

visiting the California Science Center on behalf of their children is an unreliable method of assessing learning outcomes, let alone future life chances.

The Limitations of Indicator-Based Impact Evaluation

To circumvent the need to rely exclusively on self-report data, Falk and Needham (2011) created a 'marker' to measure the Science Center experience. 'The idea was to find a learning equivalent of a radioactive tracer; something that in and of itself may or may not be highly important, but which could be considered an indicator of something greater that was meaningful' (Falk & Needham, 2011, p. 3). A 'marker' was defined as a single science concept, the understanding of which can be attributed to the California Science Center. Using the concept "homeostasis" as the marker, Falk and Needham aver that any increase in understanding of this principle amongst the L.A. public over the years can be attributed to the Science Center. The reason for selecting homeostasis is that those who visited the newly designed Science Center had the opportunity to watch a 10-minute show about the physiological process, featuring an animatronic woman named Tess and her animated sidekick Walt. The purpose of the show was to 'tangibly and engagingly teach visitors this important, but relatively poorly understood scientific concept' (Falk & Needham, 2011, p. 3). Using this 'marker', Falk and Needham hoped to provide empirical evidence that a visit to the California Science Center directly contributed to public understanding of science. In so doing, they aimed to transcend the limitations of using self-reports for impact measurement.

Using the homeostasis marker as an impact indicator falls short firstly because no valid baseline measurement was developed in order to gauge whether actual learning had occurred. Falk and Needham instead inferred a baseline from research they conducted with visitors to the Science Center in 1998. This 1998 visitor-only sample was asked to define homeostasis prior to entering the Science Center. In this earlier study, 7% of the 1998 visitor sample was deemed to have correctly defined homeostasis. This 7% figure was considered a conservative estimate of the baseline for L.A. public's understanding of homeostasis. Thus, it is inferred that 'the percentage of those in the L.A. area able to correctly identify homeostasis prior to opening of the Science Center can be assumed to have been 7% or less' (Falk & Needham, 2011, p. 8). We would challenge the use of this 1998 sample as an estimate for the baseline of the L.A. public's understanding of homeostasis for number of reasons, including: (1) the baseline sample excludes non-visitors to the California Science Center, and (2) the self-selected sample is unlikely to be representative of the wider Los Angeles population, and is certainly not a probability sample and (3) there is no evidence provided that the same standards for determining a correct definition were applied consistently and reliably across the 1998, 2000 and 2009 datasets. Indeed, the reliability of the scoring procedure for an acceptable definition of homeostasis is not demonstrated for the 1998, 2000 or 2009 studies. What were the criteria for an acceptable (i.e. correct) definition? How many different coders were involved in

making these judgments? Were the same coders used at each time point? How was reliability ensured? In methodological terms, this kind of scoring would be considered a form of content analysis (Krippendorff, 2013; Neuendorff, 2002). Good practice in content analysis requires the reporting of inter-coder reliability statistics to show the level of error present in the scoring. That is, how highly correlated are the scores of different coders if they analyze the same content independently using the same criteria? Without gathering and presenting evidence of a reliable scoring procedure, this entire outcome measure is put in doubt.

Finally, the results of the homeostasis marker do not support the narrative that the California Science Center delivered long-term positive learning impacts for the L.A. population. In 2000, 10% of respondents sampled could provide an acceptable definition of homeostasis, nearly a decade later this figure doubled to 20%. However, 75% of those who provided an acceptable definition of homeostasis in 2000 reported they had visited the Science Center; in 2009, only 61% of those offering an acceptable definition reported visiting the Science Center. Although Falk and Needham highlighted that there was a doubling in the proportion of respondents able to correctly define the marker concept, significantly fewer of these respondents had actually visited the California Science Center. This means that the reported increase in respondents providing acceptable definitions from 10% in 2000 to 20% in 2009 cannot plausibly be attributed to the influence of the Science Center. The authors' suggestion that the change over a decade in the L.A. public's understanding of the concept homeostasis provides strong evidence that the Science Center was responsible for improving public long-term science knowledge and understanding is simply mistaken. Clearly other factors are at work in this claimed increase in understanding of homeostasis.

Evaluating the Statistical Analysis

Clearly the limitations of Falk and Needham's (2011) study are many and various. We do not have the space for a full review of the statistical methods employed in the study. In brief, more sophisticated statistical tests such as multiple linear regression or generalized linear mixed models would have been more appropriate to account for the relative contribution of a series of independent variables that could have contributed to the outcomes Falk and Needham measured. Moreover, while effect sizes are reported to a limited extent, their implications are not reflected upon in the body of the article. For example, the table illustrating 'differences in [the self-reported] amount informed about science and technology based on whether respondents had visited the Science Center' (2011, p. 8) employs a *t*-test measure to compare the level of self-reported feelings of being informed about science amongst Science Center visitors and non-visitors. While there is a difference between visitors and non-visitors on this outcome variable both in 2000 and 2009, the effect sizes are remarkably small. The reported effect sizes for the difference between visitors and non-visitors was $r_{pb} = 0.18$ in 2000 and $r_{pb} = 0.17$ in 2009. These effect sizes mean

that whether or not someone visited the California Science Center only accounted for 3.24% (2000) or 2.89% (2009) of the variance in "feeling well informed" about science and technology. Given that this statistical test is merely correlational, this difference might be expected to be much greater as those who feel well informed about science may be more likely to want to visit a Science Center. Regardless, the very small level of variance in the outcome variable that is explained by whether someone visited the Science Center is not commented on at all in the paper, nor are the effect sizes for other independent variables such as educational attainment or income level provided for comparison.

Aside from the relatively unsophisticated nature of the statistical analysis, Falk and Needham (2011) tend to frame their findings of correlations between Science Center visiting patterns and self-reported knowledge as evidence of impact. For example, they frame a relationship between visiting the Science Center and self-reported knowledge in science and technology as evidence that the Science Center increases public understanding of these subjects: 'the more frequently an individual visited this Science Center, the greater their self-reported perception that they were well-informed about science' (Falk & Needham, 2011, p. 8). This quote suggests that visiting the Science Center results in feeling better informed about science and technology. However, it is equally plausible that the causal direction of this relationship could be reversed: feeling better informed about science and technology could lead people to want to visit science centers. Indeed, throughout the article, correlations are interpreted in the most favorable possible light for claiming that the Science Center is delivering positive impacts. At this juncture, a basic precept of statistical analysis bears mentioning: Correlation is not causation.

Evaluating the Study's Claims in Light of Methodological Limitations

This methodological review of Falk and Needham's (2011) attempt to measure the long-term impacts of visiting a science center is not comprehensive. However, we have identified major issues that are important for researchers to consider when conducting this kind of study in future. To conclude this article, we briefly highlight the main claims made in this study, and the associated methodological issues we have identified up to this point.

Falk and Needham set out to address the research question, 'Does visiting the California Science Center impact public science understanding, attitudes, and behaviors, and if so, in what ways?'. This research question is not effectively addressed for a number of reasons. Firstly, they claim that their results show 'the Science Center is having an impact on the L.A. community' (Falk & Needham, 2011, p. 9). Such a generalized claim cannot be upheld when there are too many methodological and theoretical limitations. The samples were unrepresentative and introduced potentially confounding variables that were not accounted for in the analysis (i.e. higher income and educational attainment). Meanwhile, the impact

measures are only based on self-report survey questions that are poorly designed. The use of self-reports to measure learning impacts was an unreliable approach, and the survey design is fraught with limitations. For example, the exclusive use of positively framed survey items clearly increases the risk of acquiescence bias.

Moreover, the homeostasis marker does not support the suggestion that the Science Center delivered long-term educational impact. In regards to theoretical considerations, Falk and Needham never thoroughly showed how they isolated and measured one learning experience (such as attending the Science Center) and how this experience interacted within a complex and multidimensional learning infrastructure. They also never specified how they controlled for the influence of other sources of learning, as highlighted in our assessment of the homeostasis marker providing evidence of other variables.

Summarizing their findings about the long-term impacts of L.A. Science Center attendance on children, Falk and Needham (2011, p. 9) state, 'Although responses of parents about their children's experiences at the Science Center were second hand and thus need to be viewed with some caution, they were overwhelmingly positive'. As discussed above, there is obvious ambiguity and unreliability in expecting different parents to judge cognitive outcomes on behalf of their children. Asking parents to provide on-the-spot assessments of their children's learning is even more prone to error than self-reported learning outcomes. Failure to follow basic principles of survey design introduced high levels of acquiescence bias, something that was unrecognized by Falk and Needham. The survey's susceptibility for acquiescence bias may offer an explanation for the overwhelmingly positive responses found for both adults self-assessments and their reporting of the learning outcomes of children.

Falk and Needham (2011, p. 10) contend that, 'The homeostasis marker allowed this research to move beyond some of the problems with self-reported data and show that a visit to this Science Center directly contributed to public understanding of science'. Although this marker was a useful attempt to circumvent the limitations of solely relying on self-report data, it was unsuccessful. No suitable baseline measurement was taken, Falk and Needham do not provide any evidence of a reliable scoring procedure used to assess whether learning had occurred, and the results from the indicator measurement did not support the authors' conclusion that the Science Center had a positive impact. Nor does this measure establish causality, merely correlations.

Conclusion

Even if you will never personally conduct a long-term impact evaluation of informal science education activities, it is valuable to be a savvy consumer of this kind of evidence as it comes to you from various sources (including a high impact peer-reviewed journal, in the present case: Journal of Research in Science Teaching). This essay is intended to serve as a reminder of the importance of following established methodological procedures. Our aim is not to introduce new

methodology here, but to issue a clarion call for researchers taking on long-term impact evaluation studies to use the hard won insights of social scientists working to improve survey and evaluation methodology. The article that is the focus of this critique is not unique in employing problematic research methods and inferences. However, the article touts its methods as an effective way of achieving the difficult task of long-term impact evaluation of informal science learning activities, a claim we challenge in this essay.

This brief review of a notable attempt to measure the long-term impacts of visiting a science center is far from comprehensive. However, we have identified important issues for researchers to consider when conducting this kind of study in future. The most plausible option for directly measuring learning outcomes is with a repeated measures design targeting the same individuals before and after visiting the Science Center (e.g. Moss, Jensen, & Gusset, 2015). Alternatively, an experimental design could be employed with a random assignment of participants to treatment and control groups. Such designs would provide a legitimate basis for drawing inferences about impact (Wagoner & Jensen, 2014). Instead, Falk and Needham (2011) employed cross-sectional surveys with first- and third-person self-reports to evaluate learning outcomes, an approach fraught with methodological limitations. Alternatives to self-report measurements include direct measurement (including open-ended data) before and after the 'intervention' of a science center visit, coupled with longer term follow-up measures including the same individuals. Longitudinal data analysis using population surveys that include both visitors and non-visitors would be an excellent (if costly) option for this research as well, but crucially the data collection would need to follow the same individuals over time to avoid the risk of sampling bias at any stage in the data collection making the results incomparable across time. There is a strong basis for these kinds of approaches in the social scientific methodological literature. This existing literature should provide the starting point for future studies of both short- and long-term informal learning impacts.

References

Bureau, U. C. (2008–2012). American fact finder. Retrieved September 10, 2014, from *Selcted Economic Characterisitcs* 2008–2012 http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

Bureau, U. C. (2000). Profile of general demographic characteristics: 2000. Retrieved September 6, 2014, from *American factfinder*: http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

Bureau, U. C. (2010). Profile of General Population and Housing Characterisitcs: 2010. Retrieved September 6, 2014, from *American factfinder*: http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

Calsyn, R. J. (1992). Acquiescence in needs assessment studies of the elderly. *The Gerontologist, 32*(2), 246–252.

Cannel, C., Miller, P., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389–437). San Francisco: Jossey-Bass.

Dawson, E., & Jensen, E. (2011). Towards a 'contextual turn' in visitor research: Evaluating visitor segmentation and identity-related motivations. *Visitor Studies, 14*(2), 127–140.

Falk, J. H., & Gillespie, K. L. (2009). Investigating the role of emotion in Science Center visitor learning. *Visitor Studies, 12*(2), 112–132.

Falk, J. H., & Needham, M. D. (2011). Measuring the impact of a science center on its community. *Journal of Research in Science Teaching, 48*(1), 1–12.

Falk, J. H., & Needham, M. D. (2013). Factors contributing to adult knowledge of science and technology. *Journal of Research in Science Teaching, 50*(4), 431–452.

Falk, J. H., & Needham, M. D. (2016). Utilizing indicator–based methods: Measuring the impact of a science center on its community. *Journal of Research in Science Teaching, 53*(1), 65–69.

Falk, J. H., & Storksdieck, M. (2005). Using the contextual model of learning to understand visitor learning from a science center exhibition. *Science Education, 89*, 744–778.

Hood, M. (1995). A view from 'outside' research on community audiences. *Visitor Studies: Theory, Research and Practice, 7*, 77–87.

Jensen, E. (2014a). Evaluating children's conservation biology learning at the zoo. *Conservation Biology*.

Jensen, E. (2014b). The problems with science communication evaluation. *Journal of Science Communication, 1*, C04.

Jensen, E., Dawson, E., & Falk, J. (2011). Dialogue and synthesis: Developing consensus in visitor research methodology. *Visitor Studies, 14*(2), 158–161.

Jensen, E., & Lister, T. (2016). Evaluating indicator-based methods of measuring long-term impacts of a science center on its community (comment). *Journal of Research in Science Teaching, 53*(1), 60–64.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Pennsylavania: SAGE Publications.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *American Jouranl of Sociology, 65*(5), 463–467.

Miller, J. D. (2001). The acquisition and retention of scientific information by American adults. In J. H. Falk, *Free-choice science education: How we learn science outside of school* (pp. 93–114). New York: NY: Teachers College Press.

Miller, J. D. (2004). Public understanding of, and attitudes toward, scientific research: What we know and what we need to know. *Understanding of Science, 13*, 273–294.

Moss, A., Jensesn, E., & Gusset, M. (2015). Evaluating the Contribution of Zoos and Aquariums to Aichi Biodiversity Target 1. *Conservation Biology, 29*(2), 537–544.

National Science Board. (2006). *Science and engineering indicators*. Washington, DC: U.S. Government Printing Office.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Cleveland State University: SAGE Publications.

St. John, M., & Perry, D. (1993). A framework for evaluation and research: Science, infrastructure, and relationships. In S. Bicknell, & G. Farmelo, *Museum visitor studies in the 90s* (pp. 59–66). London: Science Museum.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Wagoner, B., & Jensen, E. (2014). Microgenetic evaluation: Studying learning in motion. In *Yearbook of Idiographic Science: Reflexivity and Change*. Charlotte, N.C.: Information Age Publishers.