*Comment*

# Evaluating Indicator-Based Methods of 'Measuring Long-Term Impacts of a Science Center on Its Community'

Eric Allen Jensen[1] and Thomas J.P. Lister[2]

[1]*Department of Sociology, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom*
[2]*Department of Sociology, University of Warwick, Liberty Road, Coventry CV4 7AL, United Kingdom*

*Received 25 September 2015; Accepted 25 September 2015*

Abstract: This article addresses some of the challenges faced when attempting to evaluate the long-term impact of informal science learning interventions. To contribute to the methodological development of informal science learning research, we critically examine (Falk and Needham (2011) Journal of Research in Science Teaching, 48: 1–12.) study of the California Science Center's long-term impact on the Los Angeles population's understanding, attitude and interest in science. This study has been put forward as a good model of long-term impact evaluation for other researchers and informal science learning institutions to emulate. Moreover, the study's claims about the Science Center's positive impacts have been widely cited. This essay highlights the methodological limitations of Falk and Needham's innovation of using an indicator-based impact measure (a 'marker') designed to limit their reliance on self-report data, and points to more valid options for assessing long-term learning or attitudinal impacts. We recommend that future research employ more direct measurements of learning outcomes grounded in established social scientific methodology to evaluate informal science learning impacts. © 2015 Wiley Periodicals, Inc. J Res Sci Teach 53: 60–64, 2016
**Keywords:** evaluation; impact; informal science learning; methodology; public engagement with science

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. 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 learning impact evaluation, and widely cited for its conclusion that science centers are effective at achieving long-term impact. The indicator-based 'epidemiological' approach also served as the model for a 2014 international impact evaluation study focusing on science centers conducted by the same lead author[1]. We critically review one aspect of Falk and Needham's study, the use of indicators instead of direct measurement, illustrate the issues that continue to face researchers attempting this difficult yet important task.

*Correspondence to*: E. Jensen; E-mail: e.jensen@warwick.ac.uk
DOI 10.1002/tea.21297
Published online 3 December 2015 in Wiley Online Library (wileyonlinelibrary.com).

Falk and Needham's study examines the long-term impact of 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 a growing body of research around the educational value of informal science learning institutions. For decades, these institutions have made claims about their impact on the public's 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'* purports 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: 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: 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 & Gillespie, 2009; Falk & Storksdieck, 2005; Jensen, 2014b). Most existing 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 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 the methodological design used Falk and Needham's 2011 study is a good model for informal science learning researchers to adopt. This article focuses on the limitations of the indicator-based approach used to put this 'outside-in' model into practice.

## 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: 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, it is claimed 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. The purpose of the show was to 'tangibly and engagingly teach visitors this important, but relatively poorly understood scientific concept' (Falk & Needham, 2011: 3). Using this 'marker', Falk and Needham hoped to provide empirical evidence that a visit to the California Science Center directly contributed to the public's understanding of science. In so doing, they aimed to transcend the limitations traditionally associated with 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: 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, (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 all three data collection points. 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; Neuendorf, 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 the 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 the study 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.

## Conclusion

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.

## Note

[1]Last Accessed 1 May 2015 at: http://www.life.org.uk/dump/media/international-science-centre-impact-study-international-science-centre-impact-study—final-report.pdf

## References

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

Jensen, E. (2014a). Evaluating children's conservation biology learning at the zoo. Conservation Biology, 28(4), 1004–1011.

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

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

Moss, A., Jensen, E., & Gusset, M. (2015). Evaluating the contribution of zoos and aquariums to Aichi Biodiversity Target 1. Conservation Biology, 29(2), 537–544.

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

Wagoner, B., & Jensen, E. (2014). Microgenetic evaluation: Studying learning in motion. In yearbook of idiographic science: Reflexivity and change. Charlotte, NC: Information Age Publishers.