


ARTICLE

Overall Experience Rating – Measuring Visitor Response in Museums

ANDREW J. PEKARIK , JAMES B. SCHREIBER, AND NICK VISSCHER

Abstract The authors present research comparing different measures of experience quality. Using data from visitor studies at the Denver Zoo, they claim that a question that asks visitors to rate their overall experience, when used together with fully grounded five-point ordinal response scales with a category beyond Excellent, provide better results than a number of other, commonly-used scales, including Net Promoter Score. With data from the Arthur M. Sackler Gallery, they demonstrate how this measure can be used to compare and evaluate visitor responses across exhibitions.

INTRODUCTION

Performance measurement has become a standard feature of contemporary organizations. Measurement data can shape operations, guide strategy, and enforce accountability. Deciding what to measure and how to measure it are critical questions.

What should museums measure? In addition to data on attendance, finances, collections, staff, offerings, and space, there is the important matter of the visitor experience. Evaluating visitors' responses to their time in the museum is more difficult than determining the success of a transaction such as the purchase of a product or service. The outcomes of a museum visit are diverse, unpredictable, and subtle (Pekarik 2010) and are influenced not only by the opportunities provided but also by the backgrounds, knowledge levels, and even personalities of the visitors. As a result an appropriate measure needs to allow respondents to decide for themselves on what basis they make their assessment. It needs to encompass all possible outcomes and all possible subjective viewpoints. It also needs to point directly to what

it seeks to measure, namely personal experience, rather than to an indirect, future-oriented proxy such as willingness to recommend.

In the early 2000s visitor researchers at the Smithsonian Institution tried a variety of survey questions and scales in an attempt to arrive at a single item or set of items that would accurately and reliably identify the degree to which visitors were pleased with their experience in the museum. The best of these was found to be a question and a scale that came to be called the Overall Experience Rating (OER). Since then the Smithsonian has used this measure across all of its many museums both to rate visitor experience in a museum as a whole, in individual exhibitions and displays, and in public programs. Across more than a hundred studies over more than 14 years, this measure has proven to be stable and useful. Its implementation in practice can be noted in the visitor studies made available on the website of Smithsonian Organization and Audience Research (<https://soar.si.edu/reports>).

Overall Experience Rating has recently started to gain a following in other museums as well (Henriksen 2017; Visscher et al. 2017). It

Andrew J. Pekarik (andrewpekarik@gmail.com), now an independent researcher, formerly studied museum audiences at the Smithsonian Institution. James B. Schreiber (schreiberj@duq.edu) is Professor of Statistics at Duquesne University. Nick Visscher (nvisscher@denverzoo.org) is Audience Research and Evaluation Manager at Denver Zoo in Denver Colorado.

is our intention here to provide a fuller description of the measure and its use, and to offer an empirical comparison to other related scales and measures, in the hope that Overall Experience Rating can become a standard within the field.

QUALITY OF EXPERIENCE

Performance measures of audience response are part of a larger research domain generally known as Customer Satisfaction, which includes topics such as needs, desires, emotions, and expectations. The research on satisfaction spans every area of human engagement, from automobiles to zoo experiences. Because of the breadth of coverage, there is also a breadth of definitions. Cardozo’s (1965) original experimental work paved the way for the field, but multiple definitions exist and appear to be based on the characteristics of the study at hand. Giese and Cote (2000) reviewed 20 definitions spanning 30 years of research and created thematic categories. The first is *response*, which can be emotional or cognitive in nature. Then there is *focus*, which is particular to expectations, product, consumption, or experience. Finally, there is *time frame*, which can be after consumption, after choice, based on experience over time, during the purchase, and so on. Oliver (2010, 8, emphasis in the original) states that

Satisfaction is the consumer’s fulfillment response. It is a judgment that a product/service feature or the product or service itself, provided (or is providing) a *pleasurable* level of consumption-related fulfillment, including levels of under- or over-fulfillment.

This definition’s core concept - the judgment of pleasurable fulfillment, under-fulfillment and over-fulfillment - is appropriate in the museum context. Measures based directly

on this definition are generally of two types: measures of satisfaction, and measures comparing experience with expectation. Both presuppose a clear sense of what was expected (implicitly in the case of satisfaction, and explicitly in the case of comparison). Such expectation is more likely in situations where the product/service provider bears primary responsibility for the outcome of the interaction and the consumer’s role is relatively passive. Visitors’ expectations in museums, however, tend to be vague and generic because individuals have so much control over how the encounter will unfold.

We prefer a measure that directly addresses the mental/emotional state of the visitor at the moment of response without specifying or implying any prior reference. One expression of such a measure is “Quality of Experience,” a concept that has arisen in the world of digital-user performance. Quality of experience has four key dimensions (Schatz et al. 2013):

- Subjective* – it is based in a personal perspective
- User-centric* – its core concern is the user
- Holistic* – it is a comprehensive, all-inclusive response
- Multi-dimensional* – it incorporates differences in context, users, and products

These four principles encompass the personal, individual, idiosyncratic, and varied experience of a museum visitor.

WHY QUALITY OF EXPERIENCE MATTERS

Like all those committed to the public role of museums, we want visitors to have pleasurable experiences in museums. For visitors, pleasurable experiences provide affirmation that they made the right choice in choosing a

museum experience over another option, and high quality museum experiences are likely to support a feeling of personal enrichment. For museums, visitors are a key stakeholder group whose quality of experience is likely to influence their willingness to return and to provide support (Baker and Crompton 2000). At the industry level, all museums benefit when customers find their museum experiences meaningful and enjoyable.

Survey Item

We capture these dimensions in a single, straightforward survey instruction:

Please rate your overall experience with this [museum/exhibition/event/activity/etc.] [today].

This survey item meets the four dimensions of experience quality in that it is subjective, i.e., it is not an evaluation of the product but of an individual's experience. It is user-centric in that it is only focused on the user/visitor's experience, not on the goals/aims/expectations of others. It is holistic in that it requests a comprehensive assessment. It is multi-dimensional in that it does not limit itself to any particular aspect of the visitor experience.

Response Scales

For this survey item there are many potential response options and they can be divided into three categories:

Numerical scales – ungrounded numerical scales (numbers from 1 to 5, or 1 to 7, or 1 to 10, etc.); end-grounded numerical scales (with labels at the low and high ends only); and fully grounded numerical scales (with labels for each number).

Non-numerical ordinal scales with 3–10 items (a series of words that indicate rising or falling levels, such as: Good – Better – Best).¹

Disconfirmation-Confirmation scales (sets that offer comparison of present to prior states, such as: Worse than I expected – As I expected – Better than I expected.)

In the museum context these scales behave very differently, depending on the position of the item in the survey, the scale type, the number of scale items, and the labels used.

We maintain that the overall experience rating question should be placed at or near the head of the survey, in order to capture unmediated, top-of-mind responses. Positioning the question later in a survey might cause it to be influenced by memories or concepts introduced by intervening question items.

Numerical scales have a subtle flaw in that they encourage the analyst to assume that the separations between successive numbers are equidistant and that the responses form an interval variable. (Bond and Fox 2015; Wright 1977). From the use of numbers and the wrong assumption of equal distance comes the tendency to reduce the measure to a mean value. Reliance on a mean value masks the distribution, which in most museum studies is not normal. It is precisely the distribution which really matters, not the central tendency of that distribution, because, in reality, numerical scales, whether labeled or not, are ordinal categorical variables.

Disconfirmation-Confirmation scales have subtle problems that differ according to the specific type. One common type focuses on expectation: Worse than expected – As expected – Better than expected. The problem with this scale is that an individual who is asked to reflect back to a prior mental state (i.e., what was originally expected) will have been influenced in that

memory by what has taken place since then. In other words, there is a natural tendency to under-report “Less than expected,” and to over-report “As expected.” (Yuksel and Yuksel 2001). Moreover, even when a respondent reports “Better than expected,” the relative importance or degree of that difference is unknown. There is empirical evidence that performance measures are superior to confirmation-disconfirmation scales (Yuksel and Rimmington 1998).

Non-numerical ordinal scales require that full attention be given to the distribution of responses across the categories. This distribution can be greatly affected by the words used to identify points on the scale (Krosnick and Fabrigar 1997). Research has shown that these scales work best with five items (Krosnick and Presser 2010). The most commonly used 5-point labeled rating scale is Poor – Fair – Good – Very Good – Excellent. When used in museums, this scale can show distributions that are highly skewed towards the positive end of the scale. Those who have no criticisms of their experience, for example, have no reason to give a rating below the top, Excellent.² At the same time, those who are truly excited about their experience have nowhere on the scale to go past Excellent. As a result, the scale tends to better distinguish the degree of criticism (i.e., all ratings below Excellent) than it does levels of quality within the top category, Excellent. A test with various labels demonstrates that the Excellent label has a strong draw for museum visitors.³ But “Excellent” is not “Perfect,” and because Excellent is such an easy and non-critical choice, there is a need to identify a point above Excellent that will capture the responses of those who are not just satisfied and uncritical, but truly excited about their experience.

The term that we have used above Excellent is “Superior,” although research presented later in this article indicates that “Outstanding”

appears to work as well. The key point is that the respondent has been offered an option that is better than Excellent.⁴

Overall Experience Rating

The complete expression of our quality of experience survey item, Overall Experience Rating (OER) is:

Please rate your overall experience with this [museum/exhibition/event/activity/etc.][today].

- Poor
- Fair
- Good
- Excellent
- Superior

This scale can also be used to measure expectation by altering it as follows:

How do you think you will rate your overall experience at this [museum/exhibition/event/activity/etc.] today when you leave?

- Poor
- Fair
- Good
- Excellent
- Superior

HISTORY OF OER AND DISTRIBUTION OF RATINGS

The Smithsonian Institution started using OER across all exhibitions and museums starting in 2004. In a survey study of 14 Smithsonian museums in summer of 2004, the average ratings across all the museums were as follows⁵:

- 0% Poor
- 3% Fair
- 29% Good
- 49% Excellent
- 19% Superior

This distribution resembles the median ratings across 30 exhibition studies at the Smithsonian’s Arthur M. Sackler Gallery between 2004 and 2017.

- 0% Poor
- 2% Fair
- 22% Good
- 51% Excellent
- 23% Superior

The stability of this aggregate distribution masks the variability in ratings for Superior and Poor/Fair/Good (i.e., less-than-Excellent) ratings. For example, across the 30 Sackler Gallery studies, Superior ratings for individual exhibitions ranged from 10% to 52% and less-than-Excellent ratings ranged from 7% to 44%. Excellent ratings were more stable, ranging from 41% to 61%.⁶

It would seem that the Excellent category acts as an anchor when there are options both above and below it. The usefulness of OER hinges precisely on the sensitivity of those ratings above and below Excellent. Those above Excellent reveal the feelings of those who were especially excited or moved to the degree that the word “Excellent” was no longer adequate to describe the quality of their experience. At the same time, less-than-Excellent ratings record some level of criticism or hesitation, such that the rater was unwilling to mark Excellent.

Because the Excellent category is so stable, high percentages of Superior ratings tend to accompany low percentages of less-than-Excellent ratings, and the relative size of the difference can itself be understood as a type of performance measure. It would naturally be considered desirable to have Superior ratings that exceed less-than-Excellent ratings.

The very small percentage of visitors who mark Poor or Fair justifies the recoding of the five-item OER distribution to these three

categories for analysis purposes: Less-than-Excellent, Excellent, and Superior. At the same time, we need to keep Poor and Fair in the original survey item because they help to establish a clear ordinal relationship from Poor rising to Superior or Outstanding. “Superior” and “Outstanding” in themselves are not obviously stronger categories than Excellent. It is only through the force of the scale as a whole that their position beyond Excellent becomes obvious. For this reason the order of items on the scale should either start with Poor at the left (when the format is horizontal) or at the top (when the format is vertical).

COMMON USES OF OER

Comparison of Entrance and Exit Samples

The ability to gather a measure of the quality of anticipated experience across an entrance sample allows us to determine what visitors are expecting. As noted above, the survey question is revised to read *“How do you think you will rate this [museum/exhibition/event/activity/etc.] when you leave?”* In general we can say that if entrance and exit samples are surveyed in equally representative ways, but the ratings are markedly different, there is reason to believe that the audience was unexpectedly pleased or disappointed.

Eight of the thirty Sackler Gallery exit sample surveys cited above were also accompanied by entrance survey samples that asked entering visitors for their anticipated overall experience ratings. In all eight surveys Superior ratings were higher in the exit samples than in the corresponding entrance samples. The median difference was 14 percentage points. Excellent ratings were lower in six of the pairs, equal in one, and slightly higher (by 2%) in the other. The median difference was –5%.

Less-than-Excellent ratings in the exit samples were lower than in the entrance samples in four pairs, equal in one, and slightly higher (by 2%) in the two others. The median difference was –8%. This suggests that visitors to these eight exhibitions were likely to have had better experiences than expected.

OER Comparisons

We can also compare OER across museums, across exhibitions in different museums, across time, and across various other offerings, such as events, education programs, and even individual displays. Ratings can be an important factor in motivating improvement. For example, in 2014 the Freer|Sackler reviewed overall experience ratings of 21 exhibitions from the previous 10 years. The analysis revealed that on average the eleven exhibitions studied within the most recent 5 years had significantly lower Superior ratings than the ten exhibitions from the 5 years before that. In other words, the data implied that exhibitions were not getting better from a visitor perspective. As a result there was a renewed emphasis on key features associated in the data with lower OER. More recent studies suggest that this effort was successful, as Superior ratings have improved, in some cases to a remarkable degree.

OER as an Outcome Variable

We can compare OER across various visitor segments, such as first-time and repeat visitors, members and non-members, etc. We provided a small example of this in our article on Latent Class Analysis (Schreiber and Pekarik 2014, 56). Understanding who is having a better time and who is not makes it easier to determine what might be done to improve the experience for everyone.

Hopefully, in the future, as more museums adopt OER as a measure of experience quality in their museums, it will be possible to compare the results in any one museum against a reliable industry standard.

HOW BEST TO ANALYZE CATEGORICAL DEPENDENT VARIABLES LIKE OER

If we wish to understand more deeply the reasons behind ratings, we need to analyze the relationship between OER and other variables or conditions.

The three groups of the OER scale, Superior, Excellent, and Less-than-Excellent, are categorical variables (Azen and Walker 2011). There are specific analysis techniques for categorical variables that are useful depending on the research question (Schreiber and Asner-Self 2010). For example, if the question centers on predicting OER scores with independent or predictor variables, then a logistic regression for multi-category outcomes would be the best choice. A chi-square analysis would be appropriate when seeking a simple association between OER categories and other variables, for example, first-time visitors (Schreiber and Asner-Self 2011). Another approach is to examine if there are classes of visitors that exist within the data for each OER category. Latent class analysis can be used to determine the number of classes (groups) across a large number of categorical and continuous variables. (Schreiber and Pekarik 2014). For example, using variables such as OER, gender, age group, and first-time visit, it would be possible in LCA to determine how each of those variables can classify people into different groups. Finally, when the question is the difference in scores after two different versions of an exhibition, a Mann-Whitney analysis would be suitable, or for three or more versions, a Kruskal-Wallis analysis would serve.

LIMITATIONS OF OER AND HOW TO DEAL WITH THEM

Any measure can be gamed, and OER is no exception. In particular, the more narrowly and precisely targeted an audience, the easier it is to obtain a high Superior rating. For example, we find that visitors who come to a museum specifically to see a particular exhibition will tend to rate their overall experience in that exhibition more highly than visitors who came across it by accident. For this reason, it is useful to compare the OER of different sub-groups such as intentional vs. unintentional visitors. One can feel confident that an exhibition is successful when Superior ratings are high for both unintentional and intentional visitors.

Because the percentage of Superior ratings are typically around 20 percent, relatively large sample sizes are needed to confirm that the difference between two OER results are statistically significant. For example in a sample size of 100 cases, a 20% rating will have a 95% margin of error of 8%. In a sample of 300 cases, the margin of error is reduced to 5%. The need for substantial samples sizes could be a constraint in small museums without an established evaluation program. But if you collect data over time, even in relatively small samples, you will obtain a distribution of percentages in each category. These percentages will provide an overall idea of where the general percentages for your museum should be.

EMPIRICAL EVIDENCE

This section uses the 1,602 cases of the Denver Zoo Exit Survey 2017 (DZ 2017). This was a survey administered by email to visitors who were intercepted on entrance over a 4-month period (January-April, 2017) and asked to provide their email addresses for a follow-up

online survey. Surveys were sent out on the day of the visit with a 2-day follow-up reminder and a 4-day follow-up reminder for non-respondents. Overall response rate was 55%. In order to test claims made in this paper, the response scales for OER were changed randomly for subsamples as described below. Other variables included Net Promoter Score (NPS) and a confirmation-disconfirmation question on expectation.

Respondents to the Denver Zoo exit survey were randomly assigned one of five scales for the same question:⁷

Please rate the overall experience of your most recent visit to Denver Zoo.

The five scales, together with their results, are shown in Figure 1.

The first two scales had nearly identical distributions, suggesting that the two labels for the top category, “Superior” and “Outstanding”, were equivalent. The other three scales are all negatively skewed, with a higher proportion of ratings in the top category.

Earlier in this paper we proposed that the attraction of “Excellent” is that it offers a positive response without any suggestion of criticism. In the first two scales (Figures 1 and 2), where Excellent is not the top category, there is a clear choice between a category that is beyond Excellent and categories that are below Excellent. In the other scales, whether or not Excellent is included, there is a tendency for the top category to draw the highest percentage of ratings. In these cases, we believe, the top category includes both those who felt that any marking below the top category would reflect some degree of criticism, and those who felt that their experience was truly special. As a result the distribution is skewed to the high end. We can see this more clearly by considering the percentage of respondents in each of these categories who also noted that

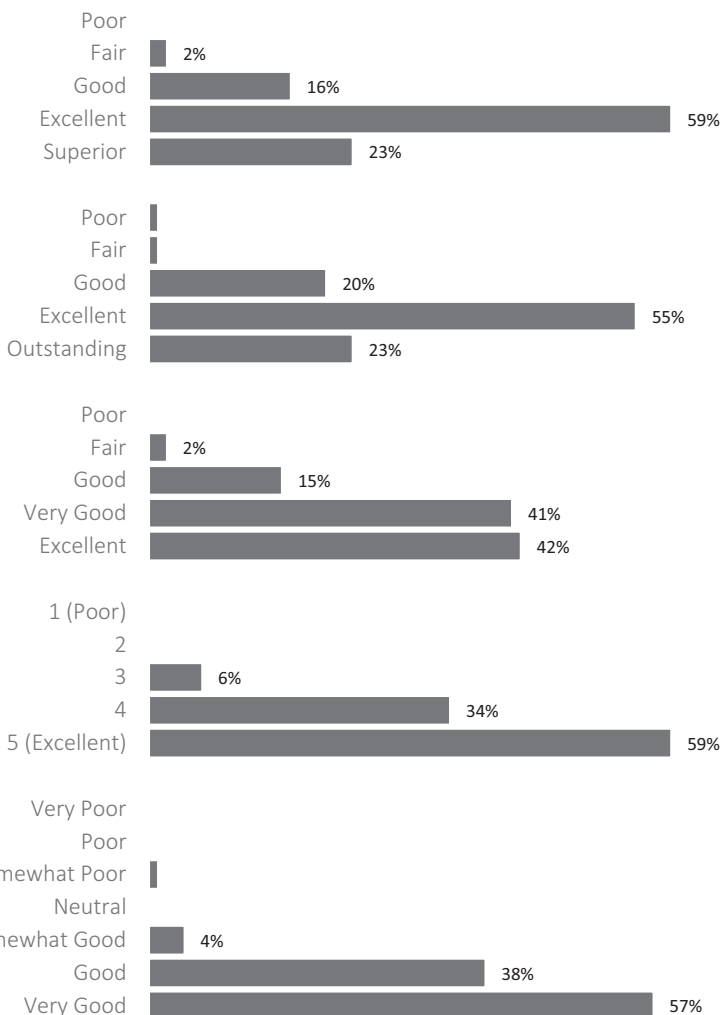


Figure 1. "Please rate the overall experience of your most recent visit to Denver Zoo"

Five Randomly Assigned Response Scales

Source: DZ 2017, N = 1,602.

their experience exceeded their expectations (see Figure 2).

Note that in the first two scales, the ones with options beyond Excellent, most respondents who chose the top category also indicated that the experience exceeded their expectations. On the other three scales roughly half of those in the top category did not feel that the experience had exceeded their expectations. Our claim is that the top categories on the first

two scales more accurately identify those who had experiences of particularly high quality.

Superiority of OER to Net Promoter Score

Recently some museums have been experimenting with the use of Net Promoter Score (NPS), a measure widely promoted in the commercial sector, to evaluate museums and exhibitions. In our opinion, this is a mistake.

Fell below or Met expectations | Exceeded expectations

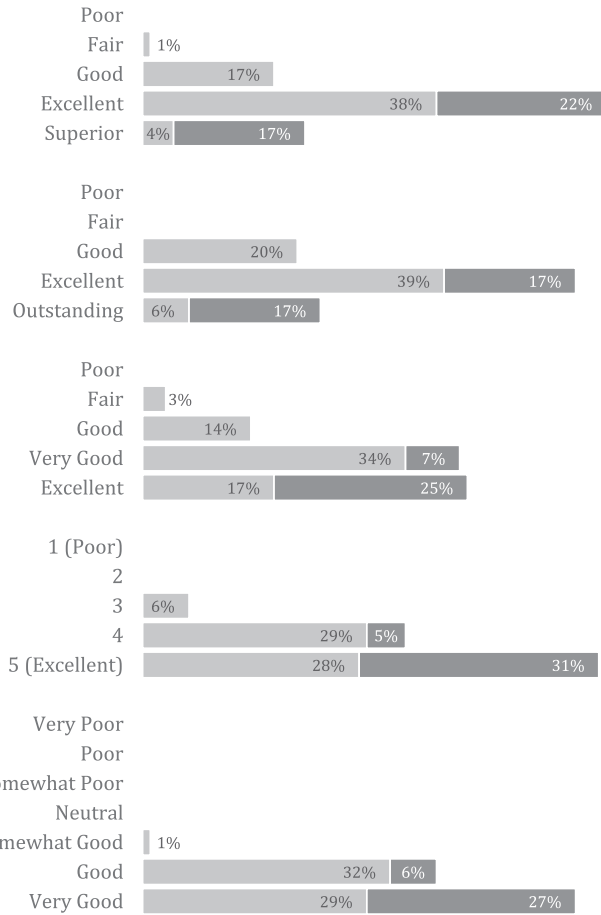


Figure 2. "Please rate the overall experience of your most recent visit to Denver Zoo"
 Five Randomly Assigned Response Scales by Expectations
 Source: DZ2017, N = 1,602.

Numerous academic articles have pointed out the deficiencies of NPS as a measure. NPS is based on a single question with an end-grounded numerical scale:

How likely is it you would recommend this [museum/exhibition/event/etc.] to a friend?



The percentages of raters in the 9 and 10 category are then classified as “Promoters”, those in categories 7 and 8 are called “Passives,” and those in categories 0 through six are called “Detractors.” A single Net Promoter Score is created by subtracting the percentage of Detractors from the percentage of Promoters.

Before illustrating some of the specific weaknesses of this measure in the museum setting, we need to point out some of the most relevant criticisms from non-museum researchers.

First, NPS was designed to be a measure of loyalty. Aside from the fact that this does not seem to be the case (Grisaffe 2007), loyalty is not a relevant concern for museums. Unlike the purchase of a car, for example, a visit to one museum does not in any way restrict or exclude the visit to another museum. In fact, the more an individual identifies as a “museum goer,” the more likely it is that the individual will visit a variety of museums.

Second, contrary to the claims of its proponents, research has shown that NPS is not a reliable indicator of an organization’s ability to grow (Keiningham et al. 2007, 2008).

Third, NPS, and the intention to recommend, as well as actual recommendations, have been shown to have little predictive value for repurchase or future business performance (Morgan and Rego 2006).

Fourth, willingness to recommend varies dramatically across different sectors and markets (Brandt 2007). While its usefulness for established, known products is limited, there is some evidence that NPS is an effective predictor of repeat purchase behaviors for new customers (Huang and Wang 2014).

Fifth, NPS, when it collapses aggregate percentages into a single score, no longer applies

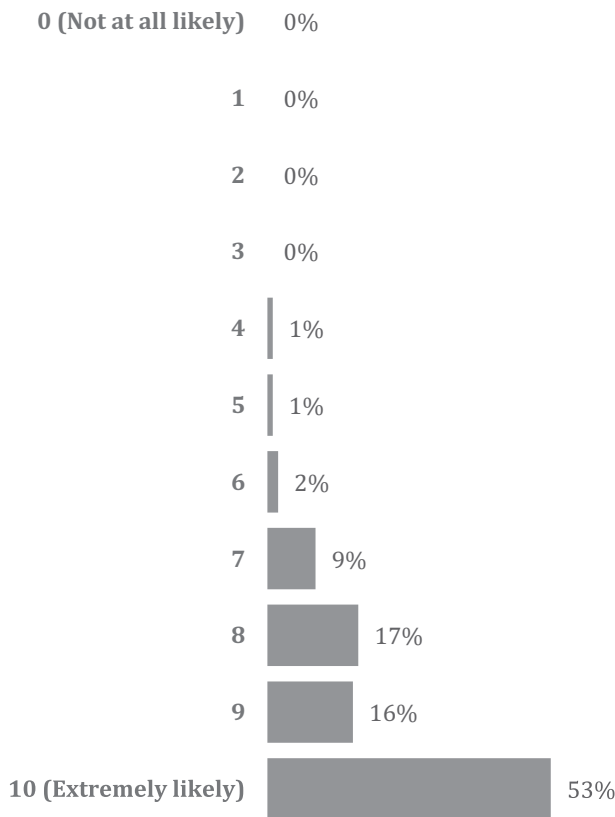


Figure 3. “Considering your most recent visit, how likely are you to recommend Denver Zoo to friends or family?” Net Promoter Score Scale DZ 2017, N = 1,615.

Table 1.

Overall experience rating by net promoter score reduced categories (DZ 2017, N = 553)

	Detractors	Passives	Promoters	Total
Less-than-excellent	32	48	25	105
Excellent	3	84	229	316
Superior/outstanding	0	6	126	132
Total	35	138	380	553

to individual respondents, and thus resists analysis. And the use of the full recommendation scale as a numerical measure for individuals has the same limitation as any end-grounded numerical scale, as mentioned above.

OER vs. NPS at Denver Zoo

The overall key to the effectiveness of OER is the fact that it produces a distribution at the high end that includes the most enthusiastic visitors.

In the Denver Zoo data the Net Promoter Score scale had over half of responses at the highest response category, as shown in Figure 3. When you add in the 16% at point 9, so-called “Promoters” (those who mark 9 or 10) comprise 69% of the whole. In such a case the Net Promoter Score, defined as the percentage marking 9 or 10 less the percentage marking 0 through 6, is 65. While this might seem like a very satisfying result, is it useful? Could a statistic that is so limited be helpful in measuring progress?

Table 1 indicates how responses to this recommendation question compared to OER.⁸ Only one-third (126 of 380, 33%) of the respondents who would be considered “Promoters” (points 9 and 10 on the NPS scale) reported a Superior experience. Nearly twice as many (229 of 380, 60%) rated their experience Excellent. Table 1 further demonstrates that although the Promoters category captured nearly all of those who rated Superior/Outstanding, it also included 73% (229 of 316) of those who rated their overall experience Excellent, and even contains one-quarter of those who gave a rating that was less than Excellent (25 of 105).

The full versions of these two measures, 5-point OER and 11-point NPS, are not measuring the same thing. OER measures experience quality and NPS measures likeliness to recommend. Although they are correlated, they only have 25% shared variance (Kendall Tau B = 0.50). In the reduced category versions of these measures, 3-point OER and 3-point NPS, the correlation is slightly weaker (Kendall Tau B = 0.49).

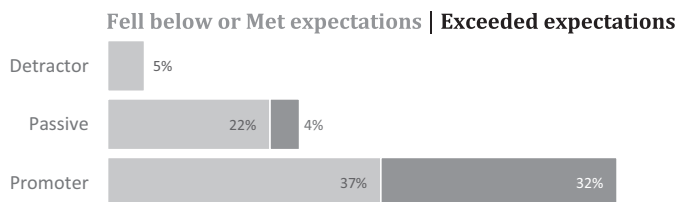


Figure 4. “Please rate the overall experience of your most recent visit to Denver Zoo”

Net Promoter Score Reduced Scale by Expectations

Source: DZ 2017, N = 1,346.

As shown in Figure 4, when we consider those within the three NPS categories who reported that the Denver Zoo experience exceeded their expectations, they comprise less than half of those in the Promoters category. This is similar to the results that we noted with the measures that did not use the Poor-Fair-Good-Excellent-Superior or Poor-Fair-Good-Excellent-Outstanding scales.

These comparisons as a whole show that NPS, the recommendation scale, is even more heavily skewed to its top category, Promoters, than the three non-OER scales seeking to measure the quality of experience.

CONCLUSION

We conclude that OER is a better predictor of enthusiastic experiences than other scales because the question addresses quality of experience broadly and the response scale includes a category beyond Excellent. Those who choose that top category are especially valuable to museums because enthusiasm is important for keeping visitors coming back and for attracting new visitors.

END

NOTES

1. It should be noted that both numerical scales and non-numerical ordinal scales can be either unipolar or bipolar. They are bipolar when positive and negative terms or values are symmetrically distributed around a neutral point in the center.
2. The high ratings visitors give to museum experiences are due in part to the tendency of many museum visitors to blame themselves when they have had an inadequate experience.
3. These points will be demonstrated with empirical data later in this article.
4. We also tested offering an option above Superior (“Optimal”), but the percentages in that category were too small to be useful in practice.

5. $N = 6,082$. See *Results of the 2004 Smithsonian-wide Survey of Museum Visitors* at <http://www.si.edu/content/opanda/docs/Rpts2004/04.10.Visitors2004.Final.pdf>. Accessed on 5 December 2016.
6. $N = 15,624$.
7. Each scale was used by at least 300 respondents.
8. The OER data cited here is the combination of the scale ending in Superior with the one ending in Outstanding.

REFERENCES

Azen, R., and C. M. Walker. 2011. *Categorical Data Analysis for the Behavioral and Social Sciences*. Abingdon, UK: Routledge.

Bond, T., and C. M. Fox. 2015. *Applying the Rasch model: Fundamental Measurement in the Human Sciences*. Abingdon, UK: Routledge.

Baker, D. A., and J. L. Crompton. 2000. “Quality, Satisfaction and Behavioral Intentions.” *Annals of Tourism Research* 27(3): 785–804.

Brandt, D. R. 2007. “For Good Measure—On the One Number You Need to Grow, One Size Doesn’t Fit All.” *Marketing Management* 16(1): 20.

Cardozo, R. N. 1965. “An Experimental Study of Customer Effort, Expectation, and Satisfaction.” *Journal of marketing research* 2(3): 244–9.

Giese, J. L., and J. A. Cote. 2000. “Defining Consumer Satisfaction.” *Academy of Marketing Science Review* 1: 1–27.

Grisaffe, D. B. 2007. “Questions About the Ultimate Question: Conceptual Considerations in Evaluating Reichheld’s Net Promoter Score (NPS).” *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior* 20: 36–53.

Henriksen, A. 2017. Normal Distribution, Ceiling Effect, and the Overall Experience Rating. [Video File]. Accessed February 12, 2018. Retrieved from <https://www.youtube.com/watch?v=IJaqTbO9ngI>

Huang, G., and H. Wang. 2014. “Improving the Predictive Validity of NPS in Customer

- Satisfaction Surveys.” In *International Conference on Cross-Cultural Design*, edited by P. L. Patrick Rau, 458–69. Cham Switzerland: Springer International Publishing.
- Keiningham, T. L., B. Cooil, T. W. Andreassen, and L. Aksoy. 2007. “A Longitudinal Examination of Net Promoter and Firm Revenue Growth.” *Journal of Marketing* 71(3): 39–51.
- Keiningham, T. L., L. Aksoy, B. Cooil, T. W. Andreassen, and L. Williams. 2008. “A Holistic Examination of Net Promoter.” *Journal of Database Marketing and Customer Strategy Management* 15(2): 79–90.
- Krosnick, J. A., and L. R. Fabrigar. 1997. “Designing Rating Scales for Effective Measurement in Surveys.” In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, de Leeuw E., C. Dippo, N. Schwarz and D. Trewin, 141–64. Hoboken, NJ: John Wiley & Sons, Inc.
- Krosnick, J. A., and S. Presser. 2010. “Question and Questionnaire Design.” In *Handbook of Survey Research*, 2nd edn, edited by P. V. Marsden and J. D. Wright, 263–314. Bingley, UK: Emerald Group Publishing Limited.
- Morgan, N. A., and L. L. Rego. 2006. “The Value of Different Customer Satisfaction and Loyalty Metrics in Predicting Business Performance.” *Marketing Science* 25(5): 426–39.
- Oliver, R. L. 2010. *Satisfaction: A Behavioral Perspective on the Consumer*. Abingdon, UK: Taylor & Francis.
- Pekarik, A. J. 2010. “From Knowing to not Knowing: Moving Beyond “Outcomes”.” *Curator: The Museum Journal* 53(1): 105–15.
- Pekarik, A. J., and J. B. Schreiber. 2012. “The Power of Expectation.” *Curator: The Museum Journal* 55(4): 487–96.
- Schatz, R., T. Hoßfeld, L. Janowski, and S. Egger. 2013. “From Packets to People: Quality of Experience as A New Measurement Challenge.” In *Data Traffic Monitoring and Analysis*, edited by E. Biersack, C. Callegari and M. Matijasevic, 219–63. Berlin Heidelberg: Springer-Verlag.
- Schreiber, J. B., and K. Asner-Self. 2010. *Educational Research*. Hoboken, NJ: Wiley Global Education.
- . 2011. *Educational Research: The Interrelationship of Questions, Sampling, Design, and Analysis*. Hoboken, NJ: John Wiley & Sons.
- Schreiber, J. B., and A. Pekarik. 2014. “Technical Note: Using Latent Class Analysis versus K-means or Hierarchical Clustering to Understand Museum Visitors.” *Curator: The Museum Journal* 57(1): 45–60.
- Visscher, N., A. Pekarik, K. DiGiacomo, and H. Ridenour. 2017. *Beyond Excellent: The Overall Experience Rating*. Session presented at the annual conference of the Visitor Studies Association, Columbus, OH.
- Wright, B. D. 1977. “Solving Measurement Problems with the Rasch Model.” *Journal of Educational Measurement* 14(2): 97–116.
- Yuksel, A., and M. Rimmington. 1998. “Customer-Satisfaction Measurement.” *Cornell Hospitality Quarterly* 39(6): 60–70.
- Yuksel, A. and F. Yuksel. 2001. “The Expectancy-Disconfirmation Paradigm: A Critique.” *Journal of Hospitality and Tourism Research* 25(2): 107–31.