EVIDENCE-BASED SURVEY DESIGN: CEILING EFFECTS ASSOCIATED WITH RESPONSE SCALES

Seung Youn (Yonnie) Chyung | Douglas Hutchinson | Jennifer A. Shamsy

Ceiling effects may negatively impact survey instruments' ability to capture survey respondents' true perspectives. This article explains the ceiling effect in surveys and explores tactics that can be applied to survey-instrument design to reduce the presence of this ceiling effect in survey data. These tactics include using more than three options in response scales, increasing the number of positive options, and using fully or even partially labeled response scale points.

INTRODUCTION

You, as a performance improvement professional, would likely use self-administered survey questionnaires to collect data and make evidence-based decisions. Your survey questionnaires may include a battery of closed-ended survey items to measure a specific performance improvement factor such as learner satisfaction, job satisfaction, organizational culture, and so forth. These closed-ended survey items include response scales that generate quantitative data to be analyzed.

When designing closed-ended survey items, you will have to make several survey-design decisions. First, the basic structure of a closed-ended survey item consists of two parts: (1) a survey-query part (i.e., what you want to ask your respondents about) and (2) a survey-response part (i.e., what you want your respondents to use to respond). You should think about how to format *the survey-query part* first—that is, should you use a question format or a statement format? Then, you need to think about how to design *the survey-response part*—that is, how do you want to capture your survey participants' responses? Table 1 shows examples.

Let's say you need to measure program participants' satisfaction levels. There are many response scales that you can use. Here are a few examples:

A. Very dissatisfied, Somewhat dissatisfied, Neutral, Somewhat satisfied, Very satisfied

B. Very dissatisfied, Somewhat dissatisfied, Somewhat satisfied, Very satisfied

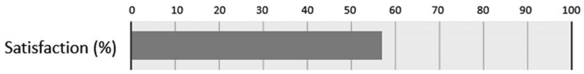
C. Very dissatisfied 1 2 3 4 5 6 7 Very satisfied

D. Not satisfied, A little bit satisfied, Moderately satisfied, Highly satisfied

E. Not satisfied 0 1 2 3 4 5 6 8 9 10 Highly satisfied

F. Very satisfied, Somewhat satisfied, Somewhat dissatisfied, Very dissatisfied

G.



These examples signal that you need to think about various design issues when selecting an appropriate response scale for your survey items. For example:

- Should you use verbal descriptive scales (e.g., A and B) or numerical rating scales (C)?
- Should you use a midpoint ("Neutral" in A) or not (no "Neutral" in B)?
- Should you use ascending order (A, B, C, D, or E) or descending order (F) of the response-scale options?
- Should you use bipolar (A, B, C, or F) or unipolar (D, E, or G) response scales?
- Should you use discrete rating scales (A through F) or continuous rating scales such as the Web-based slider

Performance Improvement, vol. 59, no. 6, July 2020

Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/pfi.21920

TABLE 1 SURVEY ITEMS DESIGNED WITH A QUESTION OR STATEMENT FORMAT

QUESTION FORMAT	STATEMENT FORMAT
How much do you feel valued at work? • Not at all • A little bit • Quite a bit • A lot	I feel valued at work. • Strongly disagree • Somewhat disagree • Neutral • Somewhat agree • Strongly agree

(G) designed to record data to a couple of decimal places (e.g., 56.23 on a scale from zero to 100)?

These seemingly simple decisions affect the type of data you will obtain. There are many studies conducted on these topics, and teams of researchers from the Organizational Performance and Workplace Learning department at Boise State University have been reviewing research articles and developing evidence-based recommendations for developing structured survey questionnaires. For example, see Chyung, Roberts, Swanson, and Hankinson (2017) on the use of a midpoint; Chyung, Barkin, and Shamsy (2018) on positively and negatively worded statements; Chyung, Swanson, Roberts, and Hankinson (2018) on incorporating continuous rating scales; and Chyung, Kennedy, and Campbell (2018) discussing ascending and descending response scale order.

This article adds to the series of these aforementioned articles on evidence-based survey design by addressing the new topic of *ceiling effects* associated with the design of response scales. The purpose of this article is twofold: (1) to explain ceiling effects associated with response scales and (2) to present research-based evidence and recommendations regarding ways to reduce ceiling effects.

CEILING EFFECTS EXPLAINED

What are ceiling effects? Several types of ceiling effects can be observed. Ceiling effects in the physical or medical sciences refer to a condition where an independent variable, such as a medical treatment, no longer produces a change on its dependent variable (i.e., the outcome). Patients will either deteriorate or no longer be able to improve because they physically or mentally cannot improve or because the measuring instrument cannot record improvement (Scherr et al., 2015). Think of a situation where your heartburn symptoms would not get any better regardless of how many heartburn relief pills you take. The effectiveness of the pills has reached the ceiling.

On the other hand, ceiling effects associated with measurements in social sciences refer to a condition where the majority of the data are close to the upper limit (Cramer & Howitt, 2004). This may occur when knowledge and ability are being tested. If your learners scored a perfect score on your test, such outcomes can be viewed positively, resulting from your effective instruction and the learners' successful learning. However, if you intended to measure improvements in their cognitive ability through an intervention, ceiling effects due to the limitations of the test instrument may result in underestimating the true effectiveness of the intervention (Scherr et al., 2015). Similarly, ceiling effects may occur when measuring gifted learners' ability with an instrument designed for non-gifted learners since the gifted learners' ratings would pile up on the top end of the instrument scale range (McBee, 2010). In these cases, learners with the same top score may still have different levels of knowledge or ability, but the measurement instrument could not capture these differences.

You may also observe ceiling effects when conducting surveys to collect facts. For example, to collect a fact about how often employees have used a job aid provided to them, suppose you asked this question: "How many times have you used the job aid?" You captured their responses with the following options: "Never," "1–2 times," "3–4 times," and "5 times or more." If most employees selected the "5 times or more" option, you would have difficulty understanding whether they meant 5 times, 18 times, or 40 times. Here, you are observing a ceiling effect.

Ceiling effects can also occur when conducting surveys to measure people's perceptions and opinions. Due to limitations in the survey instrument's sensitivity, it may be difficult to accurately measure individual respondents' true responses while distinguishing them from others' responses, resulting in little variance in the data (French et al., 2018; Taylor, 2010).

To better understand this type of ceiling effect in surveys, consider the following hypothetical scenario: You are a training manager at a manufacturing company. Recently, the company bought new equipment that technicians had never used before. Due to the complexity involved in operating the equipment and safety issues, you decided that the technicians would need to complete an instructor-led training program followed by several on-the-job coaching sessions. You administered a survey questionnaire with the technicians to measure their confidence level in using the new equipment two times—once at the end of the instructor-led training and once again at the end of the coaching sessions. The survey questionnaire included several closed-ended survey items to measure the technicians' confidence level with an *N*-point scale from low to high. You expected that the technicians would improve their confidence levels in using the new equipment after they had completed the on-the-job coaching sessions. However, when you compare survey data obtained after the instructor-led training to that obtained after the on-the-job coaching sessions, you did not see this expected improvement. The two sets of scores were both very positive, and a statistical analysis of the data did not show statistically significant improvements over time.

So, what conclusion would you draw from the results? Because the data from both surveys could not show that the on-the-job coaching sessions added value, would you discontinue the coaching sessions for future technicians? Before you do so, you might consider another possible explanation for these results—that is, on-the-job coaching sessions were indeed helpful in improving technicians' confidence levels, but the survey data might have suffered a ceiling effect. Had the survey instrument been more sensitive and capable of measuring changes in confidence, you might have received different results, avoided a ceiling effect, and reached a different conclusion.

RESEARCH ON WAYS TO REDUCE CEILING EFFECTS IN SURVEY QUESTIONNAIRES

From the foregoing hypothetical scenario, you may wonder if ceiling effects are affected by the number of options available in the response scales. In fact, researchers have investigated the occurrence of ceiling effects when using different rating scales with three, five, or more response options.

Using an Increased Number of Response Options

Research has shown that survey items using 3-point scales are more prone to ceiling effects. Scalone et al. (2013) explored the presence of the ceiling effect in a health-related quality-of-life measurement tool known as the EQ-5D. Their study asked 1,088 Italian patients with chronic hepatic diseases to describe their health across five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) using a 5-point scale and a 3-point scale. In both scales, "No problems" was the best possible option:

- 5-point scale: No problems, Slight problems, Moderate problems, Severe problems, Unable to do/extreme problems
- 3-point scale: No problems, Some problems, Unable to do/extreme problems

Patients were first given the 5-point scale, and 36.4% of them indicated "No problem" across all dimensions. These same patients were then given the survey with the 3-point scale, and 39.4% of them described themselves as having "No problems" in every single dimension. With the 5-point scale and its two additional response choices, an absolute 2.9% reduction (a relative 7.5% reduction) of ceiling effects was found.

Similarly, Feng et al. (2015) compared the same two versions of the EQ-5D and their ability to gauge the general health of UK residents. Two unique data sets were used for this study. The EQ-5D data using a 3-point scale was gathered from the 2012 Health Survey for England including 7,294 total responses. In addition, a 5-point version was used with 996 participants selected at random in England. Researchers compared the score distributions and found that offering five response options meaningfully diminished the presence of the ceiling effect. When the 3-point scale version was used, 56.2% of participants reported the best possible health state, "No problems," across every dimension. In contrast, 47.6% of participants achieved this ceiling ("No problems") when using the 5-point scale version. The researchers concluded that the 5-point scale had "greater descriptive richness" (p. 6) and improved measurement properties with a reduced ceiling effect. This allowed for more health problems to be reported and indicated that it likely had more practical usefulness than the 3-point scale.

Bharmal and Thomas (2006) also showed that expanding the number of response ratings from 3- to 5-point scales can help reduce ceiling effects in health-related quality-of-life measurements. These researchers assessed individuals' health using the EQ-5D, which measured five dimensions with 3-point options, and the SF-6D, which measured six dimensions with 5-point options. Researchers found that when the EQ-5D version was used, 47% of the 11,248 participants indicated full health ("No limitations") across all dimensions. Conversely, on the SF-6D version, only 5.8% of respondents gave a rating of full health ("No limitations"). In accordance with Feng et al.'s descriptive richness, Bharmal and Thomas (2006) concluded that the three response ratings found in the EQ-5D might not be sufficient for respondents to adequately report differences in their individual health, making a 5-point scale more appropriate.

These studies revealed the benefits of using 5-point response scales rather than 3-point response scales for reducing ceiling effects. So, you might wonder if gradual increases in the number of scale options beyond five (e.g., 7-, 9-, or 10-point scales) would correspond with a continued decrease in the ceiling effect. Research shows inconsistent conclusions. Dawes's (2008) study supported this hypothesis while Keeley et al.'s (2013) study did not. Dawes (2008) explored the differences in data characteristics between price-consciousness questionnaires that included 5-, 7-, and 10-point scales. Participants were read questions over the phone and directed to answer using a bi-directional rating scale where 1 always corresponded to "Strongly disagree" and 5, 7, or 10 (depending on the version used) equaled "Strongly agree." Results from the 5- and 7-point versions were then re-scaled so that the three versions could be compared. Dawes (2008) found that the 10-point version did help to reduce the ceiling effect by producing a 0.3-point lower average score, which was a statistically significant difference from the 5- and 7-point versions.

On the other hand, in Keeley et al.'s (2013) study, 537 university students were assigned to watch a video of a teacher exhibiting "good" or "bad" teaching behaviors. After watching their assigned video lecture, the students filled out a 28-item survey evaluating the teacher's performance, using either 5- and 7-point scales or 5- and 9-point scales. In each scale, the "Always Exhibits" rating was the highest level:

- 5-point scale: Always Exhibits, Frequently Exhibits, Sometimes Exhibits, Rarely Exhibits, Never Exhibits
- 7-point scale: Always Exhibits, Almost Always Exhibits, Frequently Exhibits, Usually Exhibits, Sometimes Exhibits, Rarely Exhibits, Never Exhibits
- 9-point scale: Always Exhibits, Almost Always Exhibits, Frequently Exhibits, Usually Exhibits, Exhibits, Sometimes Exhibits, Occasionally Exhibits, Rarely Exhibits, Never Exhibits

When they compared the responses, the researchers did not find that expanding the total number of rating options significantly increased response variability or helped to reduce ceiling effects as did Dawes's study. However, these inconsistencies in the two abovementioned studies could be explained by the differences in the response scales used and the measurement sensitivity needed for the study.

Increasing the number of response options can also be done by employing continuous rating scales such as visual analog scales (VAS). These scales allow for many more options along a continuum as compared with Likert or Likert-type discrete response scales. In a patient satisfaction/care survey of 150 Finnish orthopedic surgical patients, Voutilainen et al. (2016) found that VAS items resulted in lower average satisfaction levels, thus avoiding the ceiling effect better than Likert-scaled items. This study also found that patients, especially older patients, responded with lower satisfaction levels when they used the VAS than younger patients. This suggests that ceiling effects can differ among respondents possessing different characteristics.

Using Positively-Packed Unbalanced Response Scales

The response scales used in the previously mentioned studies are unidirectional scales; the collection of response options represents the degree of increase or decrease in measurement in one direction. Another type of response scale is a bidirectional response scale, which is usually balanced by containing the same number of positive and negative response options with or without a neutral option in the middle. Balanced bidirectional response scales such as a 5-point Likert scale (Strongly disagree, Disagree, Neutral, Agree, Strongly agree) provide respondents with two positive response options to choose from. When respondents have favorable opinions about the subject being measured, the data obtained from balanced bidirectional response scales can suffer from a lack of variance and from ceiling effects.

An alternative to a balanced bidirectional response scale is an unbalanced bidirectional response scale, which features an unequal number of positive and negative response options. This is commonly referred to as a positively-packed rating scale (Brown, 2004; Brown et al., 2009). For example, in your response scale, you might offer three or four positive options to choose from while presenting only one or two negative options. Research suggests that under the proper conditions, the use of unbalanced bidirectional/positively-packed scales can help survey designers mitigate the presence of ceiling effects.

For example, Vita et al. (2013) performed a study to examine whether the use of an unbalanced response scale could effectively reduce the ceiling effect. Their survey asked participants to rate 16 separate sessions that they attended during a 3-day academic general internal-medicine conference. Researchers collected data in 2009 and again in 2010. In 2009, 124 participants were surveyed using a paper survey with a balanced 5-point scale, and 193 individuals responded to an online survey using a 5-point unbalanced (positively packed) scale. In 2010, 718 participants were surveyed using the unbalanced scale with a small number of paper-survey respondents being excluded. The following were the scale anchors for the balanced and unbalanced scales:

- Balanced 5-point scale: 1 (Poor), 2 (Below average), 3 (Average), 4 (Above average), 5 (Outstanding)
- Unbalanced 5-point scale: 1 (Below Expectations), 2 (Average), 3 (Truly above average), 4 (Outstanding), 5 (Top 5%)

When results from 2009 and 2010 were compared, there was more variance in the data with a decrease in the

The use of an unbalanced scale produced a meaningful reduction in the presence of the ceiling effect.

number of 5 ratings from 11% to 0% in 2009 and only 1% of responses were rated a 5 in 2010. This research showed that the use of an unbalanced scale produced a meaningful reduction in the presence of the ceiling effect.

Lakin and Chaudhuri (2016) also found that positively-packed rating scales were a useful tactic for reducing ceiling effects in professional-development workshop evaluations. They asked 378 participants to complete a course evaluation after they had participated in a professional development event(s) for teaching and learning. Respondents were given one of four versions of the course evaluation that differed in the way rating scales were labeled:

- Partially-labeled 5-point scale: Low 1 2 3 4 5 High
- Partially-labeled 5-point scale: Poor 1 2 3 4 5 Excellent
- Fully-labeled positively-packed 5-point scale: 1 (Below average), 2 (Average), 3 (Above average), 4 (Well above average), 5 (Excellent)
- Fully-labeled positively-packed 5-point scale with norms: 1 (Below average), 2 (Average), 3 (Above average), 4 (Well above average, top 25%), 5 (Excellent, top 5%)

The third type, the fully-labeled positively-packed 5-point rating scale yielded a significantly lower average score, was less skewed (i.e., had reduced ceiling effects), and showed more variability in responses than the other types. Thus, the researchers recommended the use of a fully-labeled positively-packed rating scale. The fourth type (the fully-labeled positively-packed 5-point scale with norms, "top 25%" and "top 5%") did not turn out to be the better option. It was a surprising outcome because the additional wording such as "top 25%" and "top 5%" could make the anchors' meanings clearer.

It should be noted that the use of positively-packed scales is not a panacea for mitigating ceiling effects. In Masino and Lam's (2014) study, they administered the Ontario Telemedicine Network Patient Satisfaction Questionnaire to measure patient satisfaction. A convenience sample of 216 telemedicine patients completed the survey using one of three Likert scales:

- Balanced equal-interval 5-point scale: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- Positive-packed 5-point scale: Strongly Disagree, Neutral, Agree, Very Much Agree, Strongly Agree
- Positive-centered equal-interval 5-point scale: Disagree, Neutral, Agree, Very Much Agree, Strongly Agree

In their study, altering anchor labels did not prove to be a significantly effective tactic to mitigate the ceiling effect at least statistically. The percentages of participants selecting a rating of 4 or 5 when using the balanced, positive-packed, and positive-centered scales were 92.4%, 88.9%, and 90.6%, respectively. Comparing these results with those found in the research of Lakin and Chaudhuri (2016), it seems that positively-packed scales may add more benefits when the positive options are clearly distinguishable in their meanings. For example, the difference between "Well above average" and "Excellent" as found in Lakin and Chaudhuri (2016) seems clearer than the difference between "Very much agree" and "Strongly agree" as found in Masino and Lam (2014).

Labeling Response Scale Options

As reported previously, Lakin and Chaudhuri (2016) recommended using fully-labeled, positively-packed scales. Several other studies reiterate the benefits of using these fully-labeled response scales in order to reduce ceiling effects. In Garratt et al.'s study (2011), 466 patients were given one of two versions of a 24-item patient experience questionnaire (PEQ) after receiving inpatient services at a university hospital in Norway. The PEQ has traditionally used a partially-labeled 10-point scale, where only two endpoints are labeled. For example, an item inquiring about the amount of contact the patient had with nurses used the two ending anchors, "No, the nurses rarely had enough time" and "Yes, nurses always had enough time." For the purposes of this study, the PEQ was rewritten to create a fully-labeled 5-point rating scale using the anchors, "not at all," "to a small extent," "to some extent," "to a large extent," and "to a very large extent." The researchers found that the new fully-labeled 5-point iteration reduced the presence of the ceiling effect from 35.9% down to 21% as compared with the traditional partially-labeled 10-point scale. Without ceiling effects, the 5-point PEQ also resulted in lower mean scores, making it a more effective instrument for assessing patient experiences.

Even partially-labeled scales can benefit from reduced ceiling effects when compared with response options that are not labeled at all (i.e., numeric rating scales). González-Fernández et al. (2014) compared the effectiveness of a

0–10 visual analog scale (VAS) with a 0–100 general Labeled Magnitude Scale (gLMS) containing several qualitative anchors ranging from “barely detectable” to “very strong” to “the strongest imaginable sensation of any kind.” There are ongoing concerns that VASs suffer from a ceiling effect in which patients often select the highest possible response option. This is particularly problematic for practitioners because if their patients provide this skewed response, they are unable to use the tool to have patients describe an increasing degree of pain as they receive treatment. To evaluate the two scales (VAS and gLMS), González-Fernández et al. (2014) recruited 80 patients, each of whom was asked to express the pain they were experiencing using both instruments. After re-scaling of gLMS scores for direct comparison, the researchers found that “the subjects consistently rated their pain with lower scores on the gLMS than on the VAS, thus reducing the ceiling effect and allowing for rating worsening pain at the higher end of the scale” (p. 79).

SUMMARY

Ceiling effects observed in survey data can have negative impacts on your data analysis. When data is collected for summative evaluation (to assess effectiveness and accountability), skewed outcomes attributable to ceiling effects may easily be viewed favorably as evidence of a successful intervention. However, the ceiling effect may mask the true effectiveness of the intervention. Furthermore, when the intention to conduct surveys is a formative one (to make continuous improvements), skewed data under ceiling effects may prevent you from recognizing true differences among those clustered on the top (Brown & Harris, 2012; Lakin & Chaudhuri, 2016). Such data with little variance may make it difficult for you to analyze its correlations with other factors. Also, having ceiling effects would likely mean a violation of the normal data-distribution requirement, which in turn limits your statistical analyses to non-parametric methods (Šimkovic & Träuble, 2019). Therefore, the presence of the ceiling effect can affect the practical significance of the data that you receive from your survey (Kelly, 1993).

This article has reviewed research-based evidence for employing various survey design strategies and reducing ceiling effects in survey data. Research has shown that 3-point scales are prone to ceiling effects. One of the solutions to mitigate this is to increase the number of response options. Compared with 3-point scales, 5-point scales and perhaps 7- or 9-point scales are better suited for reducing ceiling effects. However, that does not mean response scales with more options are always immune from ceiling effects. Thus, the decision on how many response options to include in your survey should depend on the sensitivity that is needed in measurement.

> *Ceiling effects can be reduced by providing meaningful anchors in the response scale by fully or partially labeling them.*

When it is expected that respondents will have favorable views toward the subject being surveyed, balanced bidirectional response scales may not allow for enough positive option responses, resulting in ceiling effects and a lack of variance in survey data. To overcome this challenge, research suggests using unbalanced bidirectional response scales, aka positively-packed response scales. These scales provide more options on the positive side of the scale, making the response scale more sensitive in capturing true ratings. Research also suggests that ceiling effects can be reduced by providing meaningful anchors in the response scale by fully or partially labeling them, instead of leaving the respondents to interpret their meaning.

References

Bharmal, M., & Thomas, J. III. (2006). Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value in Health*, *9*(4), 262–271. https://doi.org/10.1111/j.1524-4733.2006.00108.x

Brown, G.T.L. (2004). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports*, *94*(3), 1015–1024. https://doi.org/10.2466/pr0.94.3.1015-1024

Brown, G.T.L., Peterson, E.R., & Irving, S.E. (2009). Beliefs that make a difference: Adaptive and maladaptive self-regulation in students’ conceptions of assessment. In D.M. McInerney, G.T.L. Brown, & G.A.D. Liem (Eds.), *Research on sociocultural influences on motivation and learning. Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 159–186). Information Age Publishing.

Brown, G., & Harris, L. (2012). Student conceptions of assessment by level of schooling: Further evidence for ecological rationality in belief systems. *Australian Journal of Educational & Developmental Psychology*, *12*, 46–59. https://files.eric.ed.gov/fulltext/EJ1002246.pdf

Chyung, S.Y., Barkin, J., & Shamsy, J. (2018). Evidence-based survey design: The use of negatively-worded items in surveys.

Performance Improvement, *57*(3), 16–25. https://doi.org/10.1002/pfi.21749

Chyung, S.Y., Kennedy, M., & Campbell, I. (2018). Evidence-based survey design: The use of ascending and descending order of Likert-type response options. *Performance Improvement*, *57*(9), 9–16. https://doi.org/10.1002/pfi.21800

Chyung, S.Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement*, *56*(10), 15–23. https://doi.org/10.1002/pfi.21727

Chyung, S.Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-based survey design: The use of continuous rating scales in surveys. *Performance Improvement Journal*, *57*(5), 38–48. https://doi.org/10.1002/pfi.21763

Cramer, D., & Howitt, D.L. (2004). *The SAGE dictionary of statistics: A practical resource for students in the social sciences*. Sage.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–104. https://doi.org/10.1177/147078530805000106

Feng, Y., Devlin, N., & Herdman, M. (2015). Assessing the health of the general population in England: How do the three-and five-level versions of EQ-5D compare? *Health and Quality of Life Outcomes*, *13*(1), 171. https://doi.org/10.1186/s12955-015-0356-8

French, B., Sycamore, N.J., McGlashan, H.L., Blanchard, C.C.V., & Holmes, N.P. (2018). Ceiling effects in the Movement Assessment Battery for Children-2 (MABC-2) suggest that non-parametric scoring methods are required. *PLoS ONE*, *13*(6): e0198426, 1–22. https://doi.org/10.1371/journal.pone.0198426

Garratt, A.M., Helgeland, J., & Gulbrandsen, P. (2011). Five-point scales outperform 10-point scales in a randomized comparison of item scaling for the Patient Experiences Questionnaire. *Journal of Clinical Epidemiology*, *64*(2), 200–207. https://doi.org/10.1016/j.jclinepi.2010.02.016

González-Fernández, M., Ghosh, N., Ellison, T., McLeod, J.C., Pelletier, C.A., & Williams, K. (2014). Moving beyond the limitations of the visual analog scale for measuring pain: Novel use of the general labeled magnitude scale in a clinical setting. *American Journal of Physical Medicine & Rehabilitation*, *93*(1), 75–81. http://doi.org/10.1097/PHM.0b013e31829e76f7

Keeley, J.W., English, T., Irons, J., & Henslee, A.M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement*, *73*(3), 440–457. https://doi.org/10.1177/0013164412475300

Kelly, M.K. (1993). A revision of the spiritual well-being scale (Publication No. 9406080). [Doctoral dissertation, The University of Nebraska-Lincoln]. ProQuest Dissertations Publishing.

Lakin, J.M., & Chaudhuri, S. (2016). Getting more out of educational workshop evaluations: Positively packing the rating scale. *Educational Research Quarterly*, *40*(1), 51–69.

Masino, C., & Lam, T.C.M. (2014). Choice of rating scale labels: Implication for minimizing patient satisfaction response ceiling effect in telemedicine surveys. *Telemedicine and e-Health*, *20*(12), 1150–1155. https://doi.org/10.1089/tmj.2013.0350

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly*, *54*(4), 314–320. https://doi.org/10.1177/0016986210379095

Scalone, L., Ciampichini, R., Fagiuoli, S., Gardini, I., Fusco, F., Gaeta, L., & Mantovani, L.G. (2013). Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Quality of Life Research*, *22*(7), 1707–1716. https://doi.org/10.1007/s11136-012-0318-0

Scherr, M., Kunz, A., Doll, A., Mutzenbach, J.S., Broussalis, E., Bergmann, H.J., & Killer-Oberpfalzer, M. (2016). Ignoring floor and ceiling effects may underestimate the effect of carotid artery stenting on cognitive performance. *Journal of Neurointerventional Surgery*, *8*(7), 747–751. http://doi.org/10.1136/neurintsurg-2014-011612

Šimkovic, M., & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PloS One*, *14*(8): e0220889, 1–47. https://doi.org/10.1371/journal.pone.0220889

Taylor, T.H. (2010). Ceiling effect. In N.J. Salkind (Ed.), *Encyclopedia of research design* (Vol. *1*, pp. 132–134). Sage. http://doi.org/10.4135/9781412961288.n44

Vita, S., Coplin, H., Feiereisel, K.B., Garten, S., Mechaber, A.J., & Estrada, C. (2013). Decreasing the ceiling effect in assessing meeting quality at an academic professional meeting. *Teaching and Learning in Medicine*, *25*(1), 47–54. https://doi.org/10.1080/10401334.2012.741543

Voutilainen, A., Pitkäaho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, *72*(4), 946–957. https://doi.org/10.1111/jan.12875

SEUNG YOUN (YONNIE) CHYUNG, Ed.D. is a professor and chair of the Department of Organizational Performance and Workplace Learning in the College of Engineering at Boise State University (http://www.boisestate.edu/OPWL/). She teaches graduate courses on Program Evaluation, Survey Design and Data Analysis, and Quantitative Research in Organizations. She runs a Workplace-Oriented Research Central (WORC) lab where teams of practitioners and researchers conduct research on various topics in the workplace learning and performance improvement context. She may be reached at ychyung@boisestate.edu

DOUGLAS HUTCHINSON is a graduate assistant for the Department of Organizational Performance and Workplace Learning in the College of Engineering at Boise State University. He joined Dr. Chyung's Workplace Oriented Research Central (WORC) team in August of 2019. He is currently studying to earn a Master of Science degree in Organizational Performance and Workplace Learning, which he expects to complete in May of 2021. He may be reached at DouglasHutchinson@u.boisestate.edu

JENNIFER A. SHAMSY, M.S., is a commercial airline pilot and adjunct online instructor. She graduated in Spring 2017 with a Master of Science degree in Organizational Performance and Workplace Learning and Workplace Instructional Design Graduate Certificate from Boise State University. She is currently an adjunct faculty member for the Organizational Performance and Workplace Learning department at Boise State University, teaching a course on Survey Design and Data Analysis. She may be reached at JenniferShamsy@u.boisestate.edu