CHAPTER 12

Scaling the Smileys: A Multicountry Investigation

Aaron Sedley, Yongwei Yang, and Joseph M. Paxton

Introduction

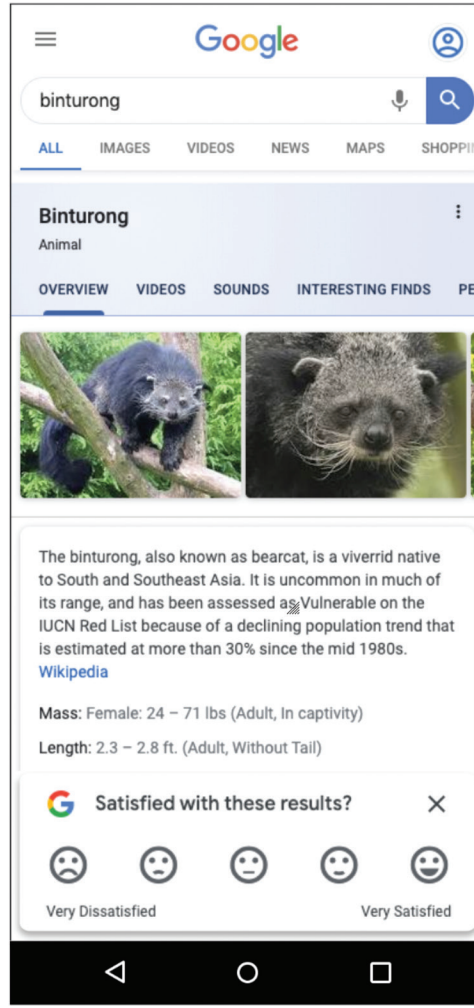
Contextual user experience (UX) surveys are brief surveys embedded in a website or mobile app (Sedley & Müller, 2016). In these surveys, emojis (e.g., smiley faces, thumbs, stars), with or without text labels, are often used as answer scales. Previous investigations in the United States found that carefully designed smiley faces may distribute fairly evenly along a numerical scale (0–100) for measuring satisfaction (Sedley, Yang, & Hutchinson, 2017). The present study investigated the scaling properties and construct meaning of smiley faces in six countries. We collected open-ended descriptions of smileys to understand construct interpretations across countries. We also assessed numeric meaning of a set of five smiley faces on a 0–100 range by presenting each face independently, as well as in context with other faces with and without endpoint text labels.

Contextual UX Surveys and Smiley Scales

Contextual UX surveys are widely used to measure attitudes and experiences "in context," that is, concurrent with actual product usage. Such contextual measurement is achieved by having the surveys triggered during or immediately after a user–product interaction. Because the survey is shown within an online product or app, it cannot occupy too much user interface (UI) space in its initial state, especially on mobile-sized screens. Failure to do so would render the survey experience overly obtrusive to the users, even to the point of hindering usage of the actual product. Fully labeled text scales often do not fit in this relatively small space. Instead, emoji-based answer scales may be used. Common smiley faces are typical emojis used for this purpose. The smartphone screenshot in Figure 12–1 provides an example.

In addition to saving space in product UIs, smiley face scales may increase survey response rates, due to the visual element being discoverable and differentiated when shown within a product and the one-click survey

Figure 12-1. Smartphone screenshot example of emoji-based answer scales



experience the design enables, compared with a two-step flow in which an invitation message precedes the actual question.

A basic smiley scale without labels also requires no translation, which may improve the fidelity and comparability of the responses in cross-cultural settings. Finally, a smiley scale may add an element of personality to the survey experience, making it more attractive and enjoyable for respondents;

however, bias potentially introduced by such a survey UI should also be considered.

Using Smileys for Contextual UX Surveys—Previous Findings in the United States

UX researchers and designers at Google have previously explored various emojis to identify a set of five smiley faces that may be consistently and quickly described by a broad range of users and reasonably differentiated for a 5-point satisfaction scale. During this process, the meanings of variants of smileys were gathered with open-ended construct association research, to ensure a happy or unhappy interpretation, rather than eliciting "dead," "angry," or other meanings. The final set of five faces is shown in Figure 12–2.

Our earlier studies found that a set of carefully selected smiley faces may possess desirable conceptual meaning and be perceived as distributed fairly evenly along a numerical scale (0–100) (Sedley et al., 2017). The interval-like scaling properties were further improved when a smiley was shown in context with the other four smileys rather than individually. The results were encouraging but limited to US respondents. With the global growth of online products and an increasing UX focus on serving users across languages and contexts, it became useful to understand the degree to which the smileys' scaling properties and construct interpretation reliably extended cross-culturally.

Scale-Point Interpretation and Properties

Survey research often uses answer scales constructed by placing a set of terms along a dimension—for example, satisfied to dissatisfied or agree to disagree. Respondents rate their attitudes or perceptions about an object, experience, or topic using these answer scales. Analyzing and interpreting such data requires that the scales behave in desirable ways. At a minimum, the scale points should function in the order as intended. Additionally, the endpoints should stretch to the ends of the intended dimension. If a midpoint is used, it should sit at the center of the dimension. Multiple scale points preferably

Figure 12-2. Smiley faces used for 5-point satisfaction scale

function in an interval manner, where the distances between adjacent scale points are equal throughout the scale. Finally, when comparisons are needed among populations (e.g., age, cultural groups), properties, such as ordinality, endpoint and midpoint locations, and scale-point distance, should be comparable across these populations.

Understanding the meaning and intensity of scale points and the specific words used in them has attracted research dating back several decades (e.g., Bartram & Yelding, 1973; Jones & Thurstone, 1955; Myers & Warner, 1968; Wildt & Mazis, 1978). To understand the meaning of these scale points, one may simply ask respondents to interpret the corresponding words or phrases. To measure their intensity, "direct rating," where respondents assign numeric values to these words or phrases, is often used (Onodera, Smith, Harkness, & Mohler, 2005). Onodera et al. (2005) also used these methods to investigate the meaning and intensity of text scale labels with US, German, and Japanese samples and suggested that bipolar symmetrical scales with a midpoint might be best for cross-national comparisons.

We adopted similar methods to investigate the meaning and scaling properties of smiley faces used in satisfaction ratings. Specifically, we explored the following research questions:

1. What do smiley faces mean conceptually?
2. Do satisfaction scales using smiley faces exhibit desirable properties in terms of ordinality, endpoint locations, midpoint location, and equal distance?
3. Do endpoint verbal labels improve these scaling properties?

Our study extended the research on scale-point meanings and properties to visual stimuli. Moreover, we tested the scale points in the context of the full answer scale, as opposed to only individually. Last but not least, we explored the performance of smiley face scales across six distinct cultural and language settings: the United States (English), Germany (German), Spain (Spanish), Brazil (Portuguese), India (English), and Japan (Japanese).

Methods

Sample Source

Data were collected via the Google Surveys platform (Sostek & Slatkin, 2018). Respondents reached by this platform were Internet users accessing online content.

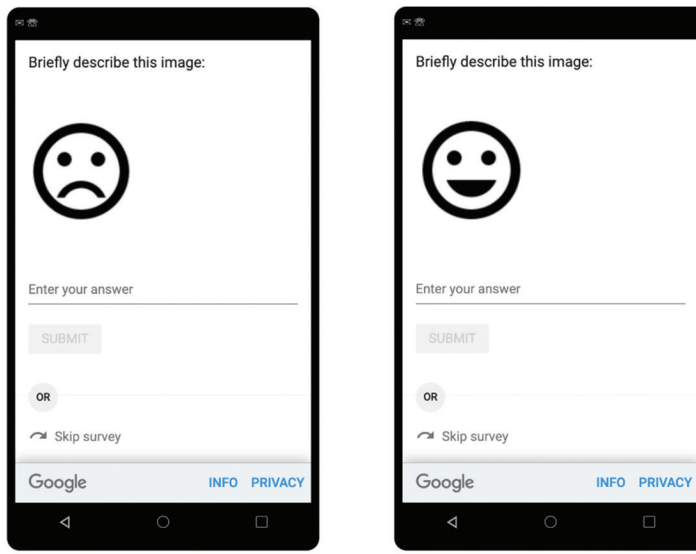
Survey Question Design

Our study tested the five smiley faces shown in Figure 12-2. Each face was tested under four conditions:

- *separate*, where only one face was presented;
- *in-scale*, where a face was highlighted within the five-face set laid horizontally from unhappiest (left) to happiest (right);
- *in-scale with "very" end labels*, similar to the *in-scale* condition with text labels "very dissatisfied" and "very satisfied" at the two ends; and
- *in-scale with "extremely" end labels*, similar to the *in-scale* condition with text labels "extremely dissatisfied" and "extremely satisfied" at the two ends.

Each respondent received one question only, asking them to either type in the meaning of a single face or assign a numeric value between 0 and 100 to the face. In the former scenario, respondents saw either the "unhappiest" or the "happiest" face, as illustrated by the smartphone screenshots in Figure 12–3. In the latter scenario, the question prompt anchored the two ends of the numeric scale as

Figure 12-3. Smartphone screenshots of meaning interpretation questions



"completely dissatisfied" and "completely satisfied," respectively. Smartphone screenshots in Figure 12–4 illustrate the respondent experience of this scenario with the *separate*, *in-scale*, and *in-scale with "very" end labels* conditions. Full question texts, endpoint labels, and their translations in Japanese, German, Spanish, and Portuguese (Brazil) can be provided upon request.

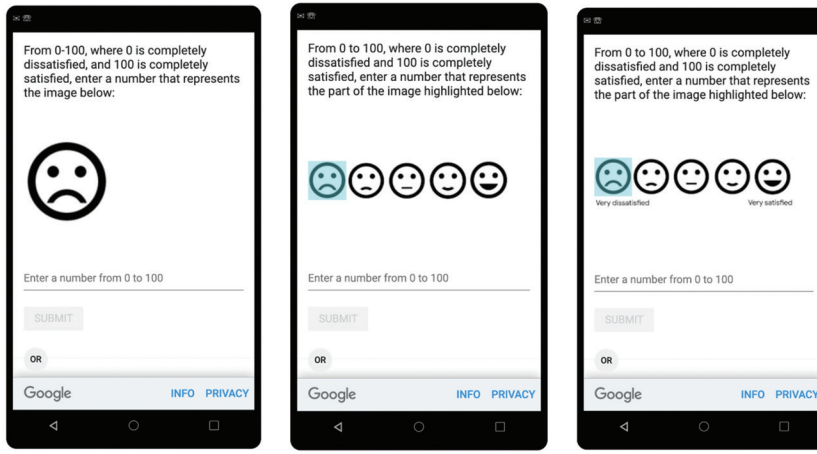
Procedures

Because respondents were asked one question only, the Google Surveys platform served a large number of surveys. Twelve 1-question surveys, two per country, were conducted to capture respondents' unaided descriptions of the smiley faces (Figure 12–3). One hundred and twenty 1-question surveys, five per country by condition combination (6 × 4), were conducted for numeric meaning of the faces (Figure 12–4).

The target sample size was 400 for each of the 12 smiley, open-ended description surveys. Target sample size for the numeric meaning surveys was 1,500. The Google Surveys platform automatically stops collecting data for a survey when the target sample size is reached. Data were collected between May and August 2019.

Respondents on the Google Surveys platform may provide suboptimal responses for various reasons. The Google Surveys platform also does not

Figure 12-4. Smartphone screenshots of numeric value questions, by condition



(a) separate (b) in-scale (c) in-scale with "very" end-labels

restrict the type of responses to an open-ended question—responses can be text or numbers of any values. Thus, for data from the numeric rating questions, we performed a series of data cleaning steps.

First, we reviewed the responses for special characters and converted them to numbers where needed. This is because respondents can input answers that, while essentially numeric, are not in Arabic numerals (e.g., 五十 in Japanese means "50") or are in multibyte format (e.g., ３５). Second, we removed the remaining non-number responses as well as those numeric responses outside the 0–100 range. Next, we reviewed the remaining responses for nonsensical values. For example, "89" is probably nonsensical as a numeric rating of the unhappiest face, whereas "6" or "4" may be nonsensical for the happiest face. To clean out such nonsensical responses, we performed a 20 percent trimming after exploring various criteria. For the directional faces (happy or unhappy), we removed 20 percent of the responses at the opposite end (e.g., 20 percent of responses in the right tail of the distribution for an unhappy face). For the neutral face, we removed 10 percent of the responses from each tail of the distribution.

The final sample sizes were 400 or slightly higher for the text interpretation surveys and ranged from 970 to 1,199 for the numeric rating surveys after data cleaning. Exact sample sizes for each survey, as well as data collection time frames, can be provided upon request.

Results

Construct Meaning

The word clouds in Figures 12–5 and 12–6 illustrate the most common associations for the happiest and unhappiest faces, respectively. (Non-English responses were first translated into English using Google Translate.) The two faces reflected the happy–sad construct consistently across the six countries. Although respondents did not naturally associate "satisfaction" or "dissatisfaction" with these faces in a survey question context, the positive–negative affective bipolarity was aligned with the measurement intent.

Scaling Properties

Figure 12–7 shows the median values of each face in each country and condition. Based on the numeric values respondents assigned, in almost all cases the smiley faces exhibited the desired ordinality—from unhappiest to happiest—and the neutral face always sat in the middle. Putting the faces in

Figure 12-5. Words and phrases associated with the happiest face 😀

context—along with other faces and in a meaningful order—improved their properties as scale points. Most noticeably, in the *in-scale* condition, the endpoints were more stretched to the extremes, and the faces were more evenly distributed, compared with the *separate* condition. Adding endpoint

Figure 12-6. Words and phrases associated with the unhappiest face ☹

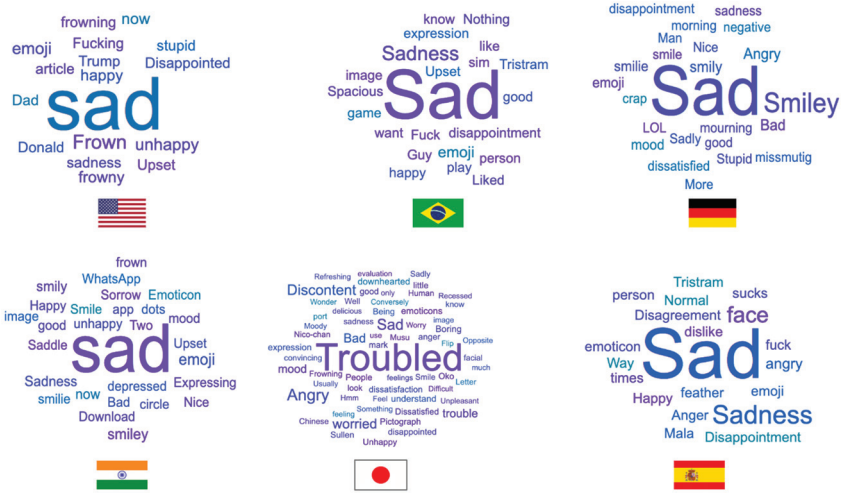
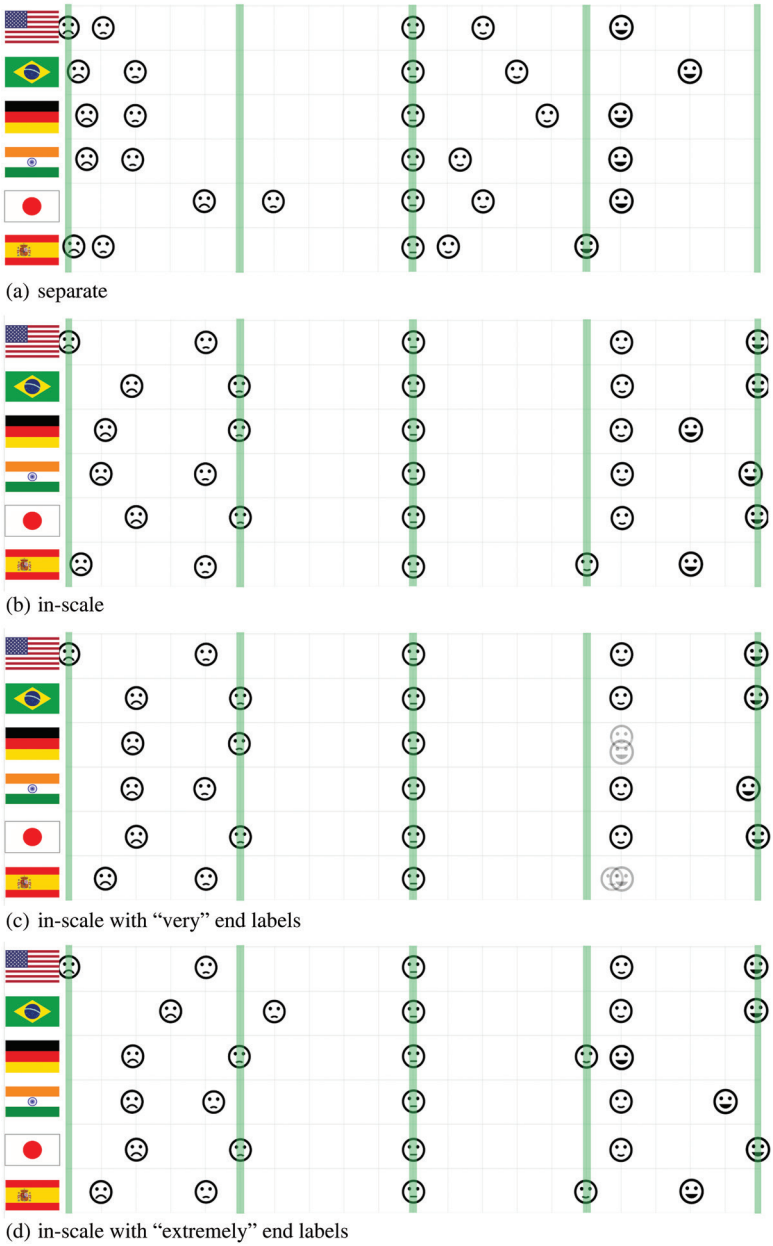


Figure 12-7. Median numeric values assigned to faces



Note: Vertical lines, from left to right, correspond to the values 0, 25, 50, 75, and 100.

Table 12-1. Deviation from ideal interval size

Condition	United States	Brazil	Germany	India	Japan	Spain
Average signed deviation from ideal interval size						
separate	–5	–3	–6	–6	–10	–7
in-scale	0	–2	–4	–2	–3	–3
in-scale with "very" end labels	0	–3	–7	–3	–3	–7
in-scale with "extremely" end labels	0	–4	–7	–4	–3	–4
Average absolute deviation from ideal interval size						
separate	15	11	13	14	10	16
in-scale	5	5	7	7	5	6
in-scale with "very" end labels	5	5	10	8	5	11
in-scale with "extremely" end labels	5	6	7	8	5	6

text labels, however, did not appear to improve the scale properties, especially in terms of endpoint locations and interval equivalence.

Table 12–1 further illustrates the findings with regard to the interval equivalence. Here we assumed that, on a 0–100 numeric scale, a 5-point answer scale's ideal interval size would be 25 (i.e., the adjacent scale points are all 25 points apart). Next, we computed the observed interval sizes using the median numeric values found for each face in each country and condition. We then computed the deviations of the observed interval sizes from the ideal of 25 in two ways, as signed or absolute differences. A zero deviation means the interval size matched the ideal. Finally, for each country and condition combination, we computed the average deviations across the four intervals. Table 12–1 presents these average deviation values. Putting faces in a scale-like context made them behave more as interval scales, whereas adding text end-labels did not bring further improvement.

Discussion

Findings from this study, regarding the construct association and scaling properties of the smiley faces, support the use of emoji-based scales for surveys across diverse countries. This is particularly encouraging for contextual UX survey applications given space constraints and response rate implications. Considering the practical difficulties and quality challenges

introduced by survey scale translation, using emoji-based scales may ease the design and implementation for multicountry surveys. Finally, from a user-centric perspective, a smiley scale may be both cognitively simpler to process and a better experience for the respondent compared with text-only scales.

However, our findings may not generalize to some scale constructs, especially those that do not possess a clear positive–negative valence that also comports with the natural happy–unhappy interpretation. The smiley faces for the endpoints may need to be further investigated in some countries (e.g., Japan, Germany), as the faces included in our study might not be interpreted with the desired extremity. Furthermore, although the text labels we used did not improve scaling properties in our study, it would still be worthwhile to test the efficacy of other text label anchors.

In a follow-up study, we are replicating the current study with groups of respondents on the Google Surveys platform who tend to be more engaged during the survey response process. Preliminary findings show that, with these respondents, the smiley face scales perform even better. Such findings highlight the importance of reducing satisficing and other less optimal response tendencies.

Our study is a first step toward understanding the validity and utility of using emoji-based answer scales. Future studies may examine various indicators of response experience and quality, such as response rate and relevant respondent engagement metrics. Last, the efficacy of emoji-based answer scales should be put to test in various real-world research contexts, including criterion-related ones, and evaluated by construct-related validity evidence.

References

Bartram, P., & Yelding, D. (1973). The development of an empirical method of selecting phrases used in verbal rating scales: A report on a recent experiment. *Journal of Marketing Research Society, 15*, 151–156.

Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. *Journal of Applied Psychology, 39*, 31–36.

Myers, J. H., & Warner, W. G. (1968). Semantic properties of selected evaluation adjectives. *Journal of Marketing Research, 5*(4), 409–412. https://doi.org/10.1177/002224376800500408

Onodera, N., Smith, T., Harkness, J., & Mohler, P. P. (2005). Methods for assessing and calibrating response scales across countries and language. *Comparative Sociology, 4*(3–4), 365–415. https://doi.org/10.1163/156913305775010106

Sedley, A., & Müller, H. (2016, May). User experience considerations for contextual product surveys on smartphones. Paper presented at 71st annual conference of the American Association for Public Opinion Research, Austin, TX. Retrieved from https://ai.google/research/pubs/pub46422/

Sedley, A., Yang, Y., & Hutchinson, H. (2017, May). To smiley, or not to smiley? Considerations and experimentation to optimize data quality and user experience for contextual product satisfaction measurement? Paper presented at the 72nd annual conference of the American Association for Public Opinion Research, New Orleans, LA. Retrieved from https://ai.google/research/pubs/pub46421

Sostek, K., & Slatkin, B. (2018, June). How Google Surveys works. Retrieved October 11, 2019, from https://services.google.com/fh/files/misc/white_paper_how_google_surveys_works.pdf

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research, 15*(2), 261–267. https://doi.org/10.2307/3151256