Journal of Usability Studies
Vol. 12, Issue 2, February 2017 pp. 73–91

User Experience Rating Scales with 7, 11, or 101 Points: Does It Matter?

James R. Lewis
Senior HF Engineer
IBM Corp.
5901 Broken Sound Parkway
Suite 514C
Boca Raton, FL 33487
USA
jimlewis@us.ibm.com

Oğuzhan Erdinç
Associate Professor
Istanbul Kultur University
Department of Industrial Engineering
Istanbul, Turkey
o.erdinc@iku.edu.tr

Abstract

There is a large body of work on the topic of the optimal number of response options to use in multipoint items. The takeaways from the literature are not completely consistent, most likely due to variation in measurement contexts (e.g., clinical, market research, psychology) and optimization criteria (e.g., reliability, validity, sensitivity, ease-of-use). There is also considerable research literature on visual analog scales (VAS), which are endpoint-anchored lines on which respondents place a mark to provide a rating. Typically, a VAS is a 10-cm line with the marked position converted to a 101-point scale (0–100).

Multipoint rating items are widely employed in user experience (UX) research. The use of the VAS, on the other hand, is relatively rare. It seems possible that the continuous structure of the VAS could offer some measurement advantages. Our objective for this study was to compare psychometric properties of individual items and multi-item questionnaires using 7- and 11-point Likert-type agreement items and the VAS in the context of UX research.

Some characteristics (e.g., means and correlations) of the VAS were different from the Likert-style (7- and 11-point items), so the VAS does not appear to be interchangeable with the Likert-style items. There were no differences in the classical psychometric properties of reliability and concurrent validity. Thus, we did not find any particular measurement advantage associated with the use of 7-point, 11-point, or VAS items. With regard to measurement properties, it doesn't seem to matter (but the literature suggests multipoint items are easier to use).

Keywords

Likert-type item, visual analog scale, VAS, number of response options, user experience

Introduction

It is common in usability testing and user experience (UX) research to collect data using multipoint rating scales. There are many questions regarding the effect of different rating scale formats on the quality of the resulting data, many of which have yet to be definitively answered, perhaps because there is a complex set of tradeoffs rather than simple answers. One of those questions is the optimal number of response options.

Usability practitioners and UX researchers currently use as few as two response options to a very large number by having participants place a mark on a 10-cm line (or use a slider control for online data collection). These visual analog scales (VAS) are usually converted into measurements that range from 0 to 100. The most common numbers of response options in standardized usability questionnaires are five (e.g., the System Usability Scale, SUS, Brooke, 1996) and seven (e.g., the Computer System Usability Questionnaire, CSUQ, Lewis, 1995), although there are popular UX instruments that use as few as three (e.g., the Software Usability Measurement Inventory, SUMI, Kirakowski, 1996) and as many as nine (e.g., the Questionnaire for User Interaction Satisfaction, QUIS, Chin, Diehl, & Norman, 1988).

When you offer two response options for a subjective experience, you can determine if a participant's experience was negative or positive, but you do not allow the expression of a neutral feeling (which you can obtain with three response options) and you do not collect any gradation of the negative or positive response (which you can obtain with a minimum of four options). The smallest number of options that includes a neutral point and gradation of negative/positive response is five. Moving beyond five allows for finer and finer gradation of the negative/positive response. It seems reasonable that increasing the number of response options should lead to improved data quality, but a review of the literature suggests that may not necessarily be the case.

Optimization Criteria

Researchers from various scientific fields have addressed the question of the optimal number of response options in different contexts with different optimization criteria, including the following:

- **Scale reliability**: Psychometric measurement of scale reliability (e.g., coefficient alpha; Alwin, 1997; Cicchetti, Showalter, & Tyrer, 1985; Jacoby & Matell, 1971; Jensen, Karoly, & Braver, 1986; Lozano, García-Cueto, & Muñiz, 2008; Matell & Jacoby, 1971; Maydeu-Olivares, Kramp, Garcia-Forero, Gallardo-Pujol, & Coffman, 2009; Preston & Colman, 2000; van Schaik & Ling, 2007)
- **Scale validity**: Psychometric measurement of some aspect of scale validity (e.g., predictive, concurrent, construct; Alwin, 1997; Briggs & Closs, 1999; Davey, Barratt, Butow, & Deeks, 2007; Jacoby & Matell, 1971; Jensen et al., 1986; Larroy, 2002; Matell & Jacoby, 1971; Maydeu-Olivares et al., 2009; Preston & Colman, 2000; Revilla, Saris, & Krosnick, 2014; van Schaik & Ling, 2007)
- **Sensitivity**: The extent to which the metric is sensitive to variation in an independent variable expected to affect the metric (Bolognese, Schnitzer, & Ehrich, 2003; Couper, Tourangeau, & Conrad, 2006; Hjermstad et al., 2011; Joyce, Zutshi, Hrubes, & Mason, 1975; Lara-Muñoz, Ponce de Leon, Feinstein, Purnte, & Wells, 2004; Larroy, 2002; Loken, Pirie, Virnig, Hinkle, & Salmon, 1987; Preston & Colman, 2000; Sauro & Dumas, 2009; van Beuningen, van der Houwen, & Moonen, 2014; van Laerhoven, van der Zaag-Loonen, & Derkx, 2004; van Schaik & Ling, 2007)
- **Ease of use**: Differences in successful use of rating scales (e.g., missing data or incorrect responses; Bolognese et al., 2003; Briggs & Closs, 1999; Couper et al., 2006; Davey et al., 2007; Funke & Reips, 2012; Hjermstad et al., 2011; van Beuningen et al., 2014; van Laerhoven et al., 2004)
- **Preference**: The number of response options that respondents prefer using (Cox, 1980; Joyce et al., 1975; Preston & Colman, 2000; van Laerhoven et al., 2004; van Schaik & Ling, 2007)

- **Structural recovery**: The extent to which continuous measures can be converted to different ordered categories and still allow recovery of the original psychometric structure (Benson, 1971; Bollen & Barb, 1981; Green & Rao, 1970, 1971)
- **Information processing**: Assessing the balance between information transmission and human processing capacity for discrimination (Cox, 1980; Hulbert, 1975; Rausch & Zehetleitner, 2014)
- **Other**: Studied unique criteria such as error relative to known values in a simulation study (Lehmann & Hulbert, 1972; Maydeu-Olivares et al., 2009), custom complex outcome metrics (Weijters, Cabooter, & Schillewaert, 2010), correlation with the magnitude of observed significance levels of statistical tests (Lewis, 1993), and frequency of marking between response options (Finstad, 2010)

Two Influential Papers

With so much research conducted over so many years in various research contexts with multiple optimization criteria, it shouldn't be surprising that there is no definitive answer. It is beyond the scope of this paper to provide a comprehensive literature review for all fields and criteria. Those who want to understand this broad context should read two broadly influential papers, one from the market research literature (Cox, 1980) and one from psychology (Preston & Colman, 2000). Following are brief summaries.

Cox (1980) published a literature review on the optimal number of response options based on published research from 1900–1980. As you might expect from such a broad literature review, the main conclusion was, "What is apparent from the extensive body of research is that there is no single number of response alternatives for a scale which is appropriate under all circumstances" (p. 418). Some of the factors that he recommended taking into account when making this decision were the following:

- **The channel capacity of the individual scale item:** The ability of a scale with two or three response options is significantly limited with regard to the amount of information it can transmit. Adding additional response options helps, but with diminishing returns.
- **The number of scaling replications:** This applies to composite scales (e.g., Likert or semantic differentials) in which responses to multiple items are combined to assess the underlying attribute (e.g., the SUS). When items are combined to form a scale, the number of response options per item becomes less important.
- **Response error:** This is, however, difficult to assess when developing measures of sentiment (e.g., perceived usability) because there is no way to know the true expected value.

Cox (1980) would not recommend a single number, but felt that because the channel capacity of items with two or three items was low but increasing the number of response options beyond nine had low marginal returns, the number of response options should be at least five and no more than nine. As Cox noted, "It is ironic that the magic number seven plus or minus two appears to be a reasonable range for the optimal number of response alternatives, despite the fact that Miller's [1956] review is not directly relevant to this question" (p. 420).

Preston and Colman (2000) conducted an experiment in which they manipulated the number of response options from two through 11 plus asking respondents to write down a number between 0 and 100 on items rating the quality of service provided by a store or restaurant familiar to the respondent. The within-subjects experimental design (n = 149 with randomized order of presentation of the items with different numbers of response options, end-anchored with *very poor* and *very good*) allowed assessment of reliability, validity, sensitivity, and respondent preference. There were no significant differences in internal consistency (measured with coefficient alpha, an estimate of scale reliability) for multi-item scales composed of the test items (ranging from 0.79 for three response options to 0.86 for 11 response options—coefficient alphas greater than 0.70 indicate acceptable scale reliability). Differences in test–retest reliability were statistically significant, but the magnitudes of the differences were small, ranging from a correlation of 0.86 for three response options to 0.94 for eight and nine options (0.92 for 11 options; .90 for 101 options). The results for various validity and sensitivity assessments were similar: either no significant difference or, where statistically significant, differences of very small magnitude. Respondents used a 101-point fill-in-the-blank scale to

rate the ease of using the different numbers of response options. Again, there were significant differences, but none were especially large, with means ranging from 74.1 (for the 101-option fill-in-the-blank item) to 83.7 (for five response options). Mean ratings exceeded 80 for three, four, five, six, seven, eight, nine, and 10 response options. Their general conclusion was "scales with small numbers of response categories yield scores that are generally less valid and less discriminating than those with six or more response categories" (p. 12).

Non-VAS Research on Optimal Number of Response Options Since 2000

Lozano et al. (2008) used simulations to explore the effect of varying correlations among items and the number of response categories per item from two to nine. The main finding that increasing the number of response options increased the reliability of the associated scales (monotonically increasing with diminished returns, except for the transition from two to three choices).

Maydeu-Olivares et al. (2009) conducted a within-subjects study with two personality questionnaires in which the questionnaire items were manipulated to provide two, three, or five response alternatives. As the number of response alternatives increased across this somewhat limited range, reliability (internal consistency) increased; there was no effect on predictive validity, and goodness of fit for item factor analysis and item response theory models decreased.

Weijters et al. (2010) studied items with four to seven response options, with and without labeling each option. They concluded that 5-point items with just endpoints labeled was best for general survey items and 7-point items were better with younger and more educated samples such as university students. These recommendations were based on complex outcome metrics for which it was difficult to distinguish practical from statistical significance.

Revilla et al. (2014) presented findings that data quality was higher with 5-point items rather than 7- or 11-point items, where quality refers to the strength of the relationship between the observed variable and the underlying construct of interest. They noted that as their quality metric declined due to increasing the number of response options, correlations with other measurements increased.

Van Beuningen et al. (2014) compared verbal label items with five response options and 11-point numerical items with the endpoints labeled. They found some distributional differences but no correlational differences with related variables. They reported more missing data for 11-point items (~2.5%) than for 5-point items (~.75%).

For the standard psychometric criteria of reliability and predictive validity, there appears to be an advantage for more response options (Lozano et al., 2008; Maydeu-Olivares et al., 2009; Revilla et al., 2014). Keeping in mind the limits of generalizability of these five studies and their varying criteria, the recommended number of response options ranged from five to nine. In this way, the research since 2000 on the optimal number of response options has been reasonably consistent with the findings of Cox (1980) and Preston and Colman (2000).

Research Including a VAS

Neither Cox (1980) nor Preston and Colman (2000) included VAS items, which were first described by Hayes and Patterson (1921). The standard VAS is a 10-cm line forming a continuous scale, the ends of which mark the minimum and maximum levels (typically labeled) of the rated attribute. Different line lengths appear to lead to similar ratings, at least, within the range of 4–10 cm (Kreindler, Levitt, Woolridge, & Lumsden, 2003). Paper and electronic versions of the VAS correlate highly (van Duinen, Rickelt, & Griez, 2008).

There have been two different applications of VAS items. One, found most frequently in the medical literature, is as a means for obtaining clinical information (e.g., self-reported amount of depression or pain) more quickly than with a more standard multi-item questionnaire (e.g., Appukuttan, Vinayagavel, & Tadepalli, 2014; de Boer et al., 2004; Hasson & Arnetz, 2005; Lee, Brown, Perantie, & Bobadilla, 2002; Zampelis, Ornstein, Franzén, & Atroshi, 2014). The other is as an alternative graphical format to use in place of Likert-type numeric scales, either for one-shot ratings or for ratings combined into multi-item scales. It is this latter application (rather than the former) that is of interest for research into the direct comparison of the number of response options.

Numerous studies have compared psychometric qualities of Likert scales and the VAS in different contexts, but, like other investigations into the number of response options, these studies have yielded contradictory findings.

For rating of chronic pain, Joyce et al. (1975) found that a VAS performed better than a Likert-type item with four response options. It was more sensitive to dosage differences, and patients indicated a slight preference for the VAS.

Jensen et al. (1986) had 75 patients rate four kinds of pain (present, least, most, and average) using six methods (four-, five-, six-, and 11-option items, 0–100 numeric fill-in-the-blank item, and VAS). All scales had similar psychometric properties. Older patients had more trouble completing the VAS. Jensen et al. recommended using the 0–100 fill-in-the-blank item due to its relative ease of administration and scoring.

Briggs and Closs (1999) found high correlations between concurrently collected five-option verbal scales and VAS. VAS was more difficult to complete for the orthopedic patients with upper extremity injuries who took part in their study.

Larroy (2002) compared a VAS and a 0–10 point numeric scale for pain assessment. The scales correlated very highly. After multiplying the ratings of the 0–10 point scale by 10, the mean difference in scale ratings was about 3. This was statistically significant, but likely of little practical significance, and of no value in selecting one format over the other.

Bolognese et al. (2003) studied differences between a VAS and five-option Likert-type item (all options labeled). They found similar results for both approaches and argued for using the Likert-style item based on ease of administration and scoring. "Although not assessed in this study, a 0–10 point discrete scale may be the most useful compromise, incorporating all positive attributes of both the visual analogue and Likert scale responses; however, this requires further study" (p. 507).

Participants in Lara-Muñoz et al. (2004) used three different items to rate the loudness of tones: VAS, a five-option verbal rating scale, and a 0–10 numeric rating (fill in the blank). There were few differences among the scales. The VAS appeared to be slightly more accurate.

Van Laerhoven et al. (2004) found that a five-option verbal-labeled Likert scale, a VAS with 10 points, and a conventional 10 cm VAS strongly correlated in measuring emotional states and quality of life of children. The children preferred the Likert-style item.

In a study with many manipulations of item format (midpoint/no midpoint, VAS feedback/no feedback, and radio buttons numbered/not numbered), Couper et al. (2006) compared a VAS with 20-point items using radio buttons or an input box. They concluded "we find no evidence for the advantages of the VAS for the types of measurement used here. Although the distributions did not differ between the VAS and the alternative approaches, the VAS suffered from higher levels of missing data, produced more breakoffs, and took longer than the other formats" (p. 243).

In Davey et al. (2007), 400 Australian women who had just visited a dedicated breast clinic completed in random order the 20-item State Trait Anxiety Inventory (STAI), a single 5-point Likert anxiety item, and a single 10 cm anxiety VAS. Both single items were significant predictors of the STAI (Likert: $r = .75$; VAS: $r = .78$). However, 11% of women incorrectly completed the VAS, limiting its usefulness.

Van Schaik and Ling (2007) included a within-subjects comparison of multi-item instruments using 7-point Likert items or 101-point VAS (0–100). Psychometric results (reliability, construct validity, sensitivity) were similar for Likert and VAS versions. A majority of participants preferred Likert over VAS (82% with $n = 103$ for a 95% adjusted-Wald binomial confidence interval ranging from 73–88%).

Sauro and Dumas (2009) compared the Single Ease Question (SEQ; a 7-point Likert-type item) with the Subjective Mental Effort Questionnaire (SMEQ; a 151-point visual scale from 0–150) for assessing perceived usability. The two approaches yielded similar results with regard to scale sensitivity.

Lee, Stone, Wakabayashi, and Tochihara (2010) reported inconclusive results of a study of many different item formats, focused on comparison with 9-point categorical scales and VAS.

"We cannot assert what is an optimal scale for the measurement of perceived thermal sensation at this time with our results" (p. 289).

Hjermstad et al. (2011) published a literature review (54 papers) of various formats for unidimensional assessment of pain intensity. They concluded that numerical rating scales (NRS, response options labeled with numbers) were generally better than verbal rating scales (VRS) or VAS. "When compared with the VAS and VRS, NRSs had better compliance in 15 of 19 studies reporting this, and were the recommended tool in 11 studies on the basis of higher compliance rates, better responsiveness and ease of use, and good applicability relative to VAS/VRS ... Overall, NRS and VAS scores corresponded, with a few exceptions of systematically higher VAS scores" (p. 1074). The most commonly used NRS (common in the assessment of pain intensity) was NRS–11 (response options from 0 to 10).

Funke and Reips (2012) published a paper entitled, "Why Semantic Differentials in Web-Based Research Should Be Made From Visual Analog Scales and Not From 5-Point Scales." The data, however, did not support this assertion (which was based on the percentage of respondents who changed their ratings while completing a survey). The difference they reported in the percentage of respondents adjusting ratings for a VAS and a five-option Likert-type item was not statistically significant.

Rausch and Zehetleitner (2014) compared a VAS with a four-option Likert-type item and reported "both visual analogue scales as well as discrete scales are reliable measures of subjective reports of global motion experience … VAS retrieves a larger amount of information than a discrete scale with four scale steps, provided that participants take their time to make the more subtle judgements" (p. 139).

In summary, a few studies have evidence supporting the use of VAS over multipoint items with regard to sensitivity (Joyce et al., 1975), respondent preference (Joyce et al., 1975), and accuracy (Lara-Muñoz et al., 2004). A few have reported better results for multipoint items than VAS with regard to completion time (Couper et al., 2006; Rausch & Zehetleitner, 2014), completion rates (Couper, 2006; Davey et al., 2007), and respondent preference (van Laerhoven et al., 2004; van Schaik & Ling, 2007). Respondents, especially in clinical settings, sometimes had more trouble physically completing the VAS than Likert-type items (Bolognese et al., 2003; Briggs & Closs, 1999; Jensen et al., 1986). The number of response options in these studies varied from four to 20, and many of them reported no significant or practical differences in psychometric properties between VAS and the various multipoint items (Bolognese et al., 2003; Couper et al., 2006; Davey et al., 2007; Larroy, 2002; Lee et al., 2010; Rausch & Zehetleitner, 2014; van Laerhoven et al., 2004; van Schaik & Ling, 2004).

Objectives of This Study

Multipoint rating items are widely used in questionnaires developed to investigate perceived usability and other aspects of the user experience. The use of the VAS, on the other hand, is relatively rare in usability studies. It is possible that the continuous structure of the VAS could offer some measurement advantages. A former disadvantage of the VAS, the need to manually score responses to items, has been eased with the introduction of tools for creating online VAS items (e.g., Marsh-Richard, Hatzis, Mathias, Venditti, & Dougherty, 2009; Reips & Funke, 2008). Despite these potential advantages, the previous literature of investigations of the VAS indicates that it might not have markedly superior psychometric properties relative to Likert-type items with enough response options to allow respondents to indicate their sentiments or judgments with reasonable precision.

Our objective for this study was to compare psychometric properties of individual items and multi-item questionnaires using 7- and 11-point Likert-type agreement items and the VAS in the context of subjective usability research. Given the broad range of previous research and multitude of criteria, we do not expect to settle these questions with one study. We do, however, hope to contribute to the scientific conversation on this topic with particular emphasis on the measurement of perceived usability.

Method

The following sections present the participants, materials, and procedures used in this study.

Participants

Fifty-eight students of a Turkish high educational institution volunteered to participate in this study. The participants were native Turkish speakers taking classes taught by the second author (Erdinç). The sample of participants included 55 males and 3 females (32 third year and 26 fourth year students) with ages ranging from 20–23 ($\bar{x} = 21.5; s = 0.86$). All students used course management software (CMS) for a variety of purposes, including

- access to course materials (96.6% reported using this feature),
- communication (17.2% reported using this feature),
- homework (51.7% reported using this feature),
- announcements (60.3% reported using this feature), and
- email (58.6% reported using this feature).

Materials and Procedure

The faculty and students are required to use the CMS, so all participants were familiar with the software. The data were collected during classes under supervision and the students were asked to respond on each scale independently from their responses on the other scales. To rate the CMS, participants used the short version of the Turkish version of the Computer System Usability Scale (T-CSUQ-SV; Erdinç & Lewis, 2013).

The T-CSUQ-SV is a standardized usability questionnaire based on the original English version, the CSUQ, which was itself based on the Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995, 2002; Sauro & Lewis, 2009, 2016). The CSUQ has been extensively used in usability research (e.g., Barak, Kastelan, & Azia, 2016; Tullis & Stetson, 2004), and the T-CSUQ-SV has been applied in recent Turkish usability research (Erdinç, Karga & Ürkmez, 2015). Table 1 shows the items from the T-CSUQ-SV (Turkish and English translation).

The T-CSUQ has been shown to have the same factor structure as the CSUQ (Erdinç & Lewis, 2013). This is important because even with careful translation of items, there is no guarantee that a translated questionnaire will have the same psychometric properties as the original version (van de Vijver & Leung, 2001). Cross-cultural research in standardized assessment of sentiment has provided some evidence that members of different cultures exhibit different levels of the extreme response tendency, although these differences do not always appear (Grimm & Church, 1999). Erdinç and Lewis (2013) found no evidence of any extreme response bias during the development of the T-CSUQ, consistent with the analysis of extreme response bias conducted on the English version with respondents from the United States (Lewis, 2002).

These questionnaires provide overall scores plus scores for three subscales: System Usefulness (SysUse), Information Quality (InfoQual), and Interface Quality (IntQual). Table 1 shows the items that go with each subscale in the T-CSUQ-SV (in Turkish and English). Note that the last item (Item 13) contributes to the overall score, but not to any subscale—a precedent established for the original PSSUQ and CSUQ.

Table 1. The T-CSUQ-SV (Short Version)

Subscale	English	Turkish
SysUse	1. Overall, I am satisfied with how easy it is to use this system.	1.Genel olarak, sistemin kullanım kolaylığından memnunum.
	2. It is simple to use this system.	2.Sistemi kullanmak basittir.
	3. I can effectively complete my work using this system.	3.Sistemi kullanarak işlerimi etkin bir şekilde yapabiliyorum.
	4. I feel comfortable using this system.	4.Sistemi rahatlıkla kullanabiliyorum.
	5. It was easy to learn to use this system.	5.Sistemi kullanmayı öğrenmem kolay oldu.
	6. I believe I became productive quickly using this system.	6.Sistemi kullanarak kısa zamanda üretken hale geldiğime inanıyorum.
InfoQual	7. The system gives error messages that clearly tell me how to fix problems.	7.Sistemin verdiği hata mesajları, problemleri nasıl gidereceğimi açıkça anlatmaktadır.
	8. The information (such as on-line help, on-screen messages and other documentation) provided with this system is clear.	8.Sistemin verdiği bilgiler (çevrim-içi yardım, ekran mesajları, diğer bilgiler, vb.) açık ve nettir.
	9. The information provided with the system is easy to understand.	9.Sistemin verdiği bilgiler kolayca anlaşılmaktadır.
IntQual	10. The interface of this system is pleasant.	10.Sistemin arayüzünü beğendim.
	11. I like using the interface of this system.	11.Sistemin arayüzünü kullanmak hoşuma gidiyor.
	12. This system has all the functions and capabilities I expect it to have.	12.Sistem, beklediğim bütün işlevlere sahiptir ve yeterlidir.
NA	13. Overall, I am satisfied with this system.	13.Genel olarak sistem tatmin edicidir.

Note. The overall score is the mean of the ratings of all 13 items; the subscale scores are the means of the ratings of their respective items.

The participants completed three alternate pen-and-paper versions of the T-CSUQ-SV, assembled into booklets with the versions presented in random order. One version used the standard item format with 7-point Likert-type items using endpoints of 1: *Strongly agree* (*Kesinlikle katılıyorum*) and 7: *Strongly disagree* (*Kesinlikle katılmıyorum*), so lower scores indicate a better experience. The other two versions replaced the 7-point Likert-type items with 11-point and VAS items but were otherwise identical. Figure 1 shows the three item formats.

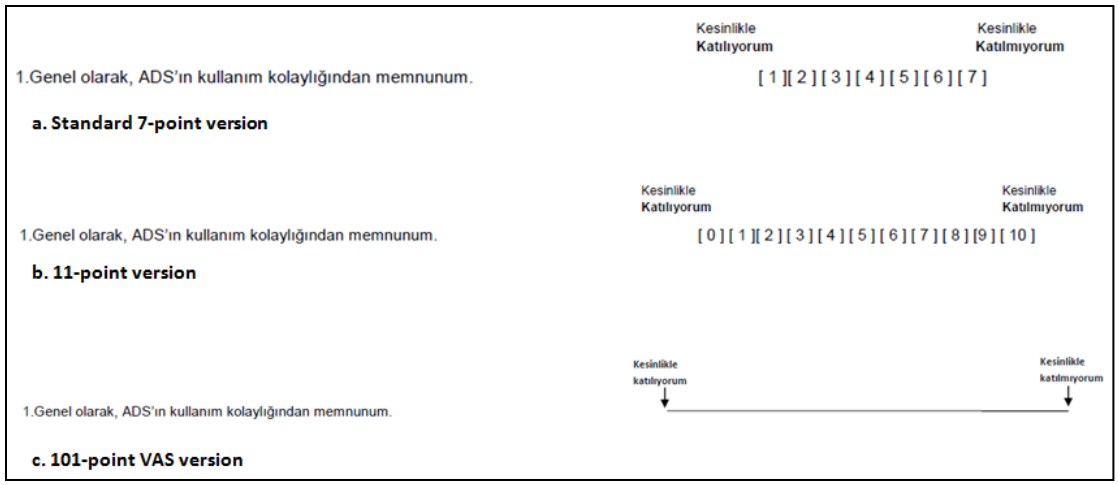

Figure 1. The three item formats (reduced as needed to fit in the space for the figure).

Results

In the following sections we present the results of this study: data conversion, item distributions, reliability, concurrent validity, means, correlations, and sensitivity.

Data Conversion

To ease comparison, all 7- and 11-point data were converted to a 0–100 (101-point) scale. For 11-point items, the conversion was to simply multiply scores by 10. For 7-point items, the conversion was to subtract 1 from the score, then multiply by 100/6 (which, for example, converts a 1 to 0 and a 7 to 100).

Item Distributions

As is typical with satisfaction ratings, Shapiro-Wilk tests of normality indicated non-normal distributions ($p < .05$) for all items except Item 13 for the 11-point and VAS formats. The distributions for the aggregated overall scales (means of the 13 items) passed the tests of normality (7-point: $p = .10$; 11-point: $p = .35$; VAS: $p = .41$). The patterns of non-normality were not consistent across the items, but, as shown in Figure 2, the distributions for the items as a function of format were consistent.

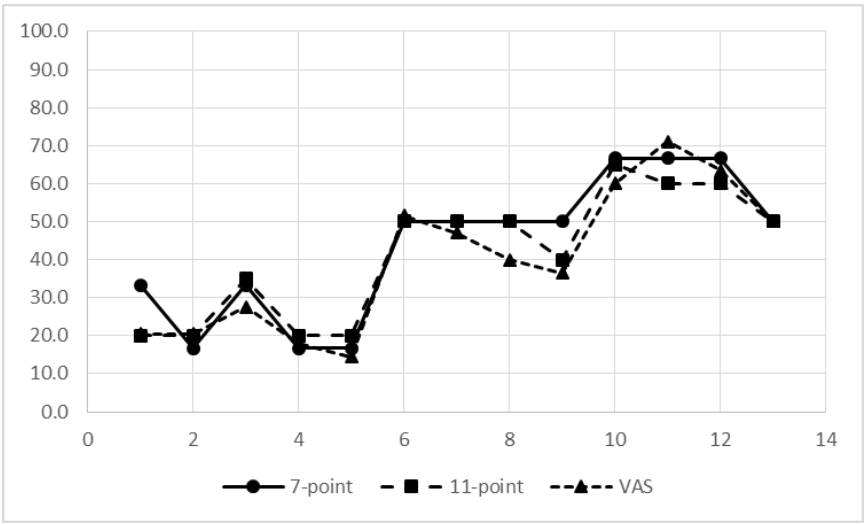

Figure 2. Item medians for the three formats.

Figure 2 shows the medians for each item. Medians around 50 indicate symmetrical distributions around the scale midpoint—about as many scores above as below the midpoint of the scale. Medians below 50 indicate scores clumped to the left of center, and those above 50 indicate the opposite pattern. For the items, the medians ranged from 14.5 to 71.0. The medians of the aggregate overall scales were close to 50 (7-point: 50.0; 11-point: 45.8; VAS: 46.0). These results show that the items under investigation covered a wide range of distributions, symmetrical and nonsymmetrical.

Correlations of the medians across formats by item were statistically significant and of very high magnitude (7-point with 11-point: $r = 0.96$; 7-point with VAS: $r = 0.96$; 11-point with VAS: $r = 0.95$—all with 11 df and $p < 0.01$). Thus, the distributions of the items appeared to be very similar for the three item formats. Note that non-normal item distributions are typically not a serious problem for most statistical analyses, which assume normality of the sampling distribution of the mean rather than normality of the underlying distribution. The distribution of the means of these types of scores tends to rapidly approach normality (Sauro & Lewis, 2016).

Reliability

Table 2 shows the scale reliabilities (coefficient alpha overall and by subscale) for each version of the T-CSUQ-SV. The typical criterion for an acceptable level of coefficient alpha for these types of scales is 0.70 (Landauer, 1997; Nunnally, 1978). The values for all three versions exceeded 0.80, with no version having an obvious advantage over another.

Table 2. Reliability Coefficients

Scale	7-point version	11-point version	VAS version
Overall	0.88	0.88	0.89
SysUse	0.87	0.88	0.86
InfoQual	0.82	0.85	0.81
IntQual	0.82	0.88	0.87

Concurrent Validity

To assess differences in concurrent validity, for each version of the T-CSUQ-SV we computed correlations for each of the first 12 items with Item 13 and averaged those correlations using the Fisher z' transformation (Sauro & Lewis, 2016). As shown in Figure 3, the mean correlation increased slightly as the number of response options increased (from 0.39 to 0.42 to 0.47), but those changes were well within the bounds of the confidence intervals. There was no obvious advantage for any version over another. All mean concurrent validities exceeded the typical minimum criterion of 0.30 (Nunnally, 1978).

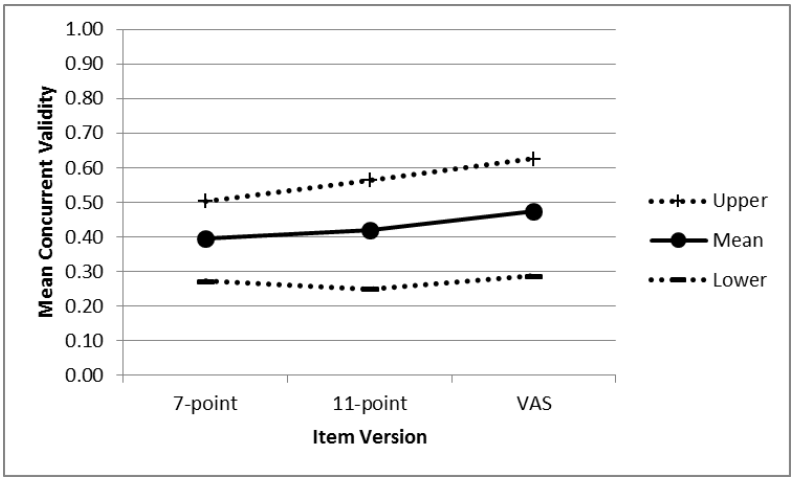

Figure 3. Concurrent validities as a function of item version with 95% confidence intervals.

Means

Figures 4 and 5 show the means for each version (with 95% confidence intervals) for, respectively, overall and subscale scores.

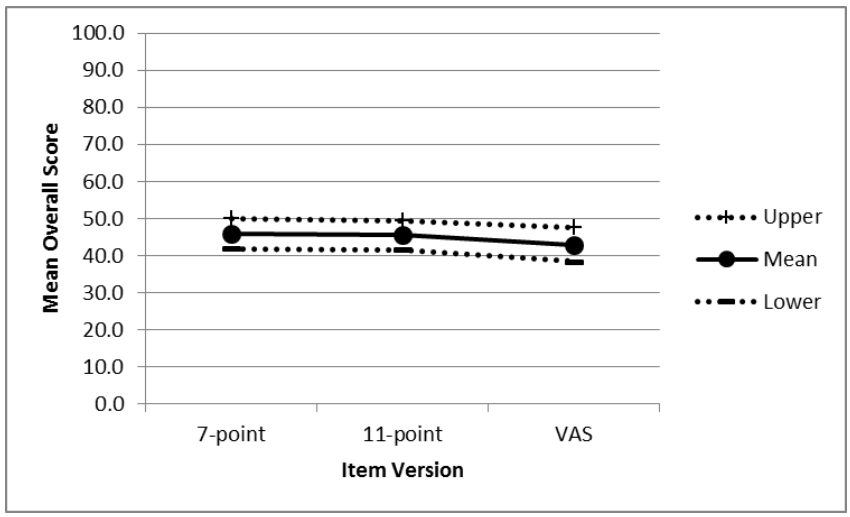

Figure 4. Means for overall scores with 95% confidence intervals.

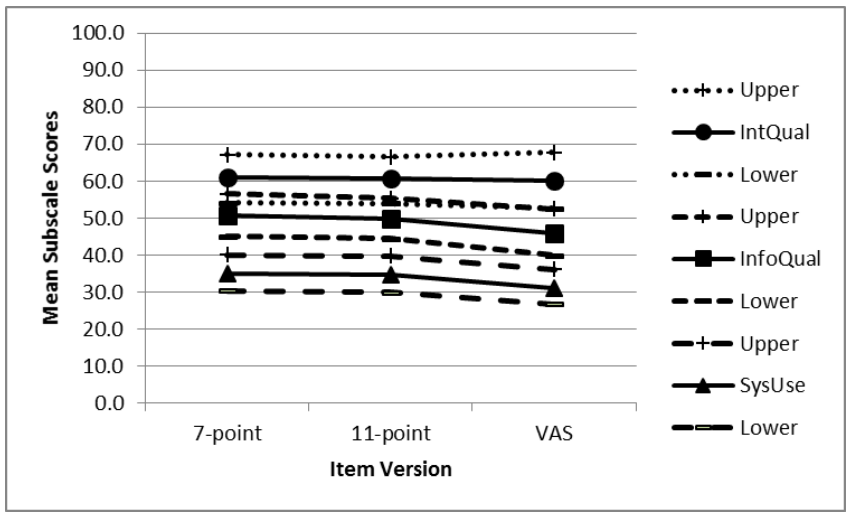

Figure 5. Subscale means with 95% confidence intervals.

Figure 4 shows a slightly lower mean for VAS (42.9) relative to 7-point (45.9) and 11-point (45.5) versions, but the confidence intervals overlapped considerably. The magnitude of the largest difference was only 3% of the range of the scale, which seems unlikely to be noticeable. Figure 5 shows that the patterns for the subscales were similar to the overall results, with a maximum difference between means of 3.9 for SysUse, 4.6 for InfoQual, and 0.9 for IntQual. Paired *t*-tests indicated statistically significant differences between VAS and the Likert-type items overall (7-point: $t(57) = 2.5$, $p = 0.02$; 11-point: $t(57) = 2.2$, $p = 0.03$) and for the SysUse (7-point: $t(57) = 2.4$, $p = 0.02$; 11-point: $t(57) = 2.4$, $p = 0.02$) and InfoQual subscales (7-point: $t(57) = 2.5$, $p = 0.02$; 11-point: $t(57) = 1.8$, $p = 0.07$). An ANOVA conducted on the results shown in Figure 5 indicated a significant main effect of Item Version ($F(1.6, 94.7) = 3.9$, $p = 0.03$—using Greenhouse-Geisser adjusted degrees of freedom due to a significant Mauchly test of sphericity), a highly significant effect of Subscale ($F(1.8, 104.7) =$

43.3, $p < 0.0001$), and no significant interaction between the two ($F(3.1, 179.0) = 0.90$, $p = 0.45$). Because the T-CSUQ-SV assesses a sentiment (perceived usability) there is no known true score against which to compare these results, so although there are some statistically significant differences, there does not appear to be an obvious advantage for one version over another.

Correlations

Figure 6 shows the mean correlations (with 95% confidence intervals), averaged across the 13 items for each pair of versions (again, using the Fisher z' transformation). Examination of the confidence intervals shows that the mean correlation between 7- and 11-point items was significantly higher than those between the Likert-style and VAS items ($p < .05$). The mean correlation between 7- and 11-point items was 0.83 (95% confidence interval ranging from 0.78 to 0.87), while their correlations with the VAS items were, respectively, 0.73 and 0.74 (95% confidence interval ranging from 0.68 to 0.78 for both correlations). This illustrates another difference in the behavior of the items, but does not indicate any particular advantage.

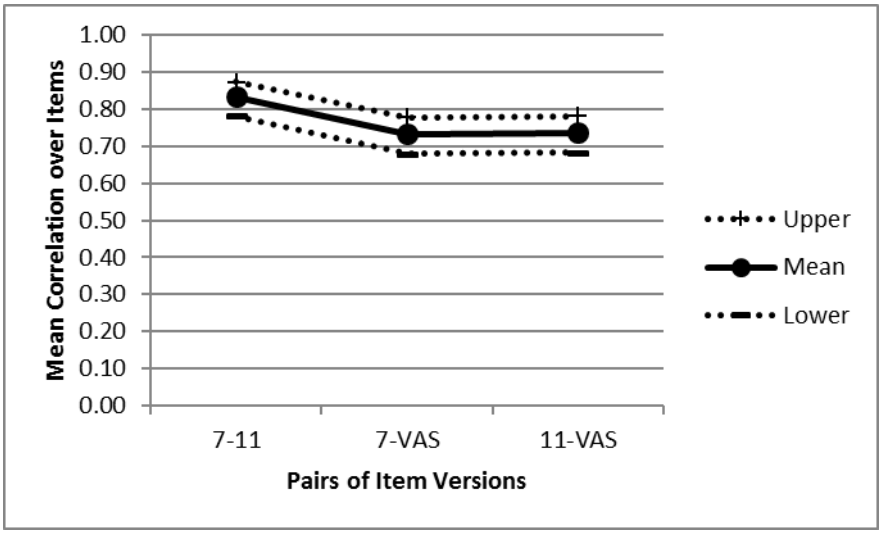

Figure 6. Mean correlations with 95% confidence intervals for the three versions.

Sensitivity

Three of the content management system components that participants rated—homework, announcements, and email—had about equal distributions. The results of a series of independent samples *t*-tests demonstrated that there were no significant differences for these variables regardless of the version of the T-CSUQ-SV used in the analysis.

Figure 7 shows the results of splitting the sample into groups based on the mean of their overall scores for all three versions of the questionnaire. The research question here was not whether there would be a significant difference between the groups—that is assured by the way the sample was split. The result of interest is whether one of the versions was more sensitive to this manipulation than the others. As the figure shows, this was not the case: The mean differences for the versions ranged from 25 for the 11-point version (95% confidence interval ranging from 20 to 30) to 27 for the 7-point version (95% confidence interval ranging from 22-32) to 30 for the VAS (95% confidence interval ranging from 25 to 35). Inspection of the confidence intervals reveals no significant difference in the magnitude of the mean differences, and therefore no evidence of a difference in sensitivity.

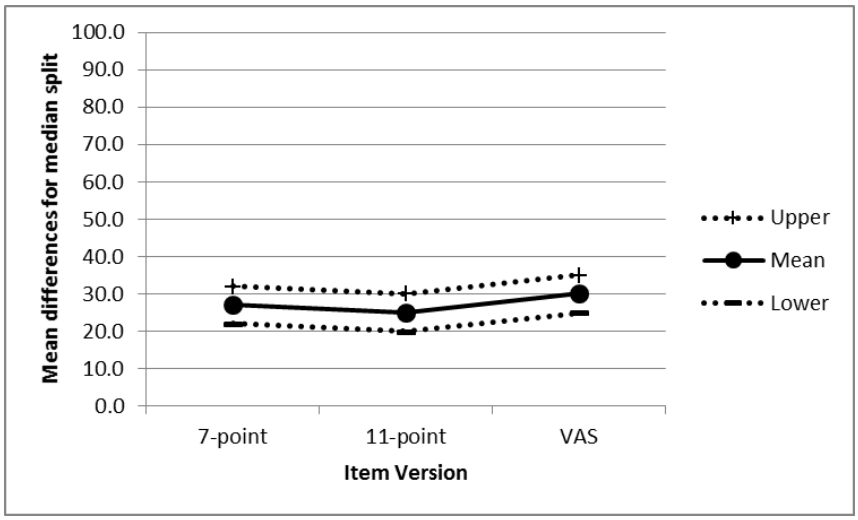

Figure 7. Mean differences of median split with 95% confidence intervals for the three versions.

The preceding analyses were for composite measurements that were means of 13 items per version. Table 3 shows the breakdown of those composites for each individual item from each version. The 95% confidence intervals in Figure 5 had widths of about ±5; for Table 3 the 95% confidence intervals were roughly ±10. This is a consequence of the increased variability for individual relative to composite measurements using these types of items, and leads to the same conclusion—there were no significant differences in the magnitude of the mean differences, and therefore no evidence of a difference in sensitivity among the versions.

Table 3. Mean Differences of Median Split for Each Item of Each Version

Item	7-point version	11-point version	VAS version
1	26	29	27
2	25	23	24
3	29	27	31
4	32	16	22
5	16	17	20
6	31	28	34
7	20	26	31
8	26	22	32
9	34	28	30
10	38	34	36
11	35	30	33
12	21	18	39
13	30	24	37

Discussion

There is a large body of work on the topic of the optimal number of response options to use in multipoint items. The takeaways from the literature are not completely consistent, most likely due to variation in measurement contexts (e.g., clinical, market research, psychology, user experience research) and optimization criteria (e.g., reliability, validity, sensitivity, ease-of-use, preference, structural recovery, and information theory). Although no single number of

response options is clearly the best, the literature generally supports the use of five to nine options.

Researchers have also studied the properties of VAS items which, strictly speaking, are not multipoint items but, in practice, are usually treated as having 101 points (ranging from 0 to 100). As with other multipoint items, the findings from the literature are not consistent. Some research has reported advantages for the VAS over other multipoint items with regard to sensitivity, respondent preference, and accuracy. A few have reported advantages for multipoint items' completion time, completion rate, respondent preference, and ease of use (especially in clinical settings where patients have difficulty using their upper bodies). For the fundamental psychometric properties of reliability, validity, and sensitivity, most studies have found no significant differences between VAS and multipoint items.

Researchers who study the user experience, either in the field or in the lab, often use multipoint items to collect ratings of the user experience. We suspect that many of them wonder if they would get a better measurement using a VAS instead because marking a position on a line seems like it should allow a more fine-grained measurement and one more likely to have an interval- or ratio-level of measurement. As we found in our literature review, despite its appeal, there is little evidence that using a VAS leads to better measurement than a multipoint item with at least five response options. The objective of our research was to compare ratings collected with three different versions of a standardized usability questionnaire, the T-CSUQ-SV, with those versions differing only in the format of the items (the 7-point item used in the standard version of the questionnaire, an 11-point version, and a VAS version).

We converted all ratings to a common 0–100 point scale. For the fundamental psychometric properties of reliability, concurrent validity, and sensitivity, there were no significant differences among the three versions. There were slight differences in the magnitudes of the means for the overall rating and questionnaire subscales, with very small and nonsignificant differences between the 7- and 11-point versions, and statistically significant (but still small) differences between the multipoint and VAS versions. This was mirrored in the examination of the correlations between the different versions for which all correlations were highly significant, but the correlation between 7- and 11-point versions was slightly (and significantly) higher than those between the multipoint and VAS versions.

The key takeaway is that no item version seemed to have an advantage over the others. There were no significant differences in reliability, concurrent validity, or sensitivity. There were some small but significant differences between the multipoint and VAS formats for means and correlations. These differences, however, do not promote the use of one format over the other. They do indicate that researchers should not expect VAS and multipoint versions of the same items to produce the same values—they will likely be close to one another, but with some difference.

Recommendations

As we stated in the introduction, we do not expect this one additional study to settle the question of the optimal number of response options. We had a number of limits to generalization due to the design of the study, including

- the use of relatively young Turkish-speaking students;
- the use of one standardized usability questionnaire (T-CSUQ-SV);
- a focus on the measurement of the perceived user experience rather than a sensory attribute such as perceived pain;
- the use of bipolar rather than unipolar items;
- the use of standard multipoint and VAS item formats rather than other formats, for example, Kunin (1955) Smiley scales;
- ratings of one CMS application; and
- a focus on the key psychometric properties of reliability, validity, and sensitivity.

Another limitation is the use of paper-and-pencil T-CSUQ forms assembled into booklets. Even though they were supervised during the study, it is possible that some participants might have flipped pages to refer back to previous ratings in an attempt to be consistent. The significant

difference in magnitude between VAS and the multipoint scales suggest that either this did not happen or participants were not able to accurately map multipoint ratings onto the VAS line. Regardless, it would be valuable for future researchers to use online surveys rather than paper-and-pencil to definitely restrict the ability of participants to refer back to previous ratings.

We encourage other user experience researchers to conduct and report similar work with other populations, questionnaires, contexts of use, and optimization criteria.

The key takeaway for researchers is to not worry too much about the number of response options in their research—to be comfortable using multipoint items rather than VAS. It is, of course, fine to use the VAS if desired, but there does not appear to be any overwhelming advantage in its use, and its use can be problematic with regard to ease of use for certain user groups.

Conclusion

Some characteristics (e.g., means and correlations) of the VAS were different from the Likert-style (7- and 11-point items), so the VAS does not appear to be interchangeable with the Likert-style items. There were no differences in the classical psychometric properties of reliability and concurrent validity. Within the limits of generalization imposed by the design of this study, there did not appear to be any particular measurement advantage associated with the use of 7-point, 11-point, or VAS items. The research literature, however, does indicate some usability issues associated with VAS, making the use of multipoint items more appealing.

Tips for Usability Practitioners

We offer the following tips to practitioners:

- Do not worry excessively about the number of response options.
- The literature indicates that it's acceptable to use from five to nine response options, with the most common choice being seven. In our current research, we found no differences between ratings from 7- and 11-point scales.
- It might seem appealing to use a VAS in place of a multipoint item, but keep in mind that in the current research the VAS did not offer any clear measurement advantage, and the literature indicates that VAS may be more difficult than multipoint items for some respondents to use.

Acknowledgements

We sincerely thank the participants who provided the data for this research. We also wish to express our appreciation to the anonymous reviewers for their guidance in crafting this paper.

References

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research*, *25*(3), 318–340.

Appukuttan, D., Vinayagavel, M., & Tadepalli, A. (2014). Utility and validity of a single-item visual analog scale for measuring dental anxiety in clinical practice. *Journal of Oral Science*, *56*(2), 151–156.

Barak, M., Kastelan, I., & Azia, Z. (2016). Exploring aspects of self-regulated learning among engineering students learning digital system design in the FPGA environment-methodology and findings. In R. Szewczyk, I. Kastelan, M. Temerinac, M. Barak, & V. Sruk (Eds.), *Embedded engineering education* (pp. 139–160). Berlin, Germany: Springer Verlag Berlin.

Benson, P. H. (1971). How many scales and how many categories shall we use in consumer research? A comment. *Journal of Marketing*, *35*, 59–61.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, *46*, 232–239.

Bolognese, J. A., Schnitzer, T. J., & Ehrich, E. W. (2003). Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. *OsteoArthritis and Cartilage*, *11*, 499–507.

Briggs, M., & Closs, J. S. (1999). A descriptive study of the use of visual analogue scales and verbal rating scales for the assessment of postoperative pain in orthopedic patients. *Journal of Pain and Symptom Management*, *18*(6), 438–446.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London, UK: Taylor & Francis.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, *9*(1), 31–36.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In *Proceedings of CHI 1988* (pp. 213–218). Washington, DC: ACM.

Couper, M. P., Tourangeau, R., & Conrad, F. G. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review*, *24*(2), 227–245.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 407–422.

Davey, H. M., Barratt, A. L., Butow, P. N., & Deeks, J. J. (2007). A one-item question with a Likert or visual analog scale adequately measured current anxiety. *Journal of Clinical Epidemiology*, *60*, 356–360.

de Boer, A. G. E. M., van Lanschot, J. J. B., Stalmeier, P. F. M., van Sandick, J. W., Hulscher, J. B. F., de Haes, J. C. J. M., & Sprangers, M. A. G. (2004). Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research*, *13*(2), 311–320.

Erdinç O., Karga H., Ürkmez A. (2015). User satisfaction and components of perceived usability for a course management software. In *ICOVACS 2015* (pp. 340–347). Istanbul, Turkey: Marmara University.

Erdinç, O., & Lewis, J. R. (2013). Psychometric evaluation of the T-CSUQ: The Turkish version of the Computer System Usability Questionnaire. *International Journal of Human-Computer Interaction*, *29*, 319–326.

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, *5*(3), 104–110.

Funke, F., & Reips, U. (2012). Why semantic differentials in Web-based research should be made from visual analog scales and not from 5-point scales. *Field Methods*, *24*(3), 310–327.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery: How many scales and response categories to use? *Journal of Marketing*, *34*, 33–39.

Green, P. E., & Rao, V. R. (1971). A rejoinder to "How many scales and how many categories shall we use in consumer research? A comment." *Journal of Marketing*, *35*, 61–62.

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*, 415–441.

Hasson, D., & Arnetz, B. B. (2005). Validation and findings comparing VAS vs. Likert scales for psychosocial measurements. *International Electronic Journal of Health Education*, *8*, 178–192.

Hayes, M. H., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*, 98–99.

Hjermstad, M. J., Fayers, P. M., Haugen, D. F., Caraceni, A., Hanks, G. W., Loge, J. H., Fainsinger, R., Aass, N., & Kaasa, S. (2011). Studies comparing numerical rating scales,

verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. *Journal of Pain and Symptom Management*, *41*(6), 1073–1093.

Hulbert, J. (1975). Information processing capacity and attitude measurement. *Journal of Marketing Research*, *12*, 104–106.

Jacoby, J., & Matell, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research*, *8*, 495–500.

Jensen, M. P., Karoly, P., & Braver, S. (1986). The measurement of clinical pain intensity: A comparison of six methods. *Pain*, *27*, 117–126.

Joyce, C. R. B., Zutshi, D. W., Hrubes, V., & Mason, R. M. (1975). Comparison of fixed interval and visual analogue scales for rating chronic pain. *European Journal of Clinical Pharmacology*, *8*, 415–420.

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 169–178). London, UK: Taylor & Francis. The online SUMI is available at www.ucc.ie/hfrg/questionnaires/sumi/index.html.

Kreindler, D., Levitt, A., Woolridge, N., & Lumsden, C. (2003). Portable mood mapping: The validity and reliability of analog scale displays for mood assessment via hand-held computer. *Psychiatry Research*, *120*, 165–177.

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology*, *8*, 65–77.

Landauer, T. K. (1997). Behavioral research methods in human–computer interaction. In Helander, M., Landauer, T. K., & Prabhu, P. (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 203-227). Amsterdam, Netherlands: Elsevier.

Lara-Muñoz, C., Ponce de Leon, S., Feinstein, A. R., Purnte, A., & Wells, C. K. (2004). Comparison of three rating scales for measuring subjective phenomena in clinical research. I. Use of experimentally controlled auditory stimuli. *Archives of Medical Research*, *35*, 43–48.

Larroy, C. (2002). Comparing visual-analog and numeric scales for assessing menstrual pain. *Behavioral Medicine*, *27*, 179–181.

Lee, J. W., Brown, S., Perantie, D. C., & Bobadilla, L. (2002). A comparison of single-item visual analog scales with a multi-item Likert-type scale for assessment of cocaine craving in persons with bipolar disorder. *Addictive Disorders & Their Treatment*, *1*(4), 140–142.

Lee, J., Stone, E. A., Wakabayashi, H., & Tochihara, Y. (2010). Issues in combining the categorical and visual analog scale for the assessment of perceived thermal sensation: Methodological and conceptual considerations. *Applied Ergonomics*, *41*, 282–290.

Lehmann, D. R., & Hulbert, J. (1972). Are three-point scales always good enough? *Journal of Marketing Research*, *9*, 444–446.

Lewis, J. R. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction*, *5*(4), 383–392.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*(3 & 4), 463–488.

Loken, B., Pirie, P., Virnig, K. A., Hinkle, R., & Salmon, C. T. (1987). The use of 0-10 scales in telephone surveys. *Journal of the Market Research Society*, *29*(3), 353–362.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73–79.

Marsh-Richard, D. M., Hatzis, E. S., Mathias, C. W., Venditti, N., & Dougherty, D. M. (2009). *Behavior Research Methods*, *41*(1), 99–106.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternative for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternative in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, *41*(2), 295–308.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.

Rausch, M., & Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and Cognition*, *28*, 126–140.

Reips, U., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, *43*(1), 73–97.

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of CHI 2009* (pp. 1599–1608). Boston, MA: ACM.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of CHI 2009* (pp. 1609–1618). Boston, MA: ACM.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Cambridge, MA: Morgan-Kaufmann.

Tullis, T. S., & Stetson, J. N. (2004). *A comparison of questionnaires for assessing website usability*. Paper presented at the Usability Professionals Association Annual Conference, June. UPA, Minneapolis, MN.

van Beuningen, J., van der Houwen, K., & Moonen, L. (2014). *Measuring well-being: An analysis of different response scales* (Discussion Paper 2014 03). The Hague: Statistics Netherlands.

van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality*, *69*, 1007–1031.

van Duinen, M., Rickelt, J., & Griez, E. (2008). Validation of the electronic Visual Analogue Scale of Anxiety. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *32*, 1045–1047.

van Laerhoven, H., van der Zaag-Loonen, H. J., & Derkx, B. H. F. (2004). A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatrica*, *3*, 830–835.

van Schaik, P., & Ling, J. (2007). Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction*, *14*(1), Article 4, 1–35.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). *The effect of rating scale format on response styles: The number of response categories and response category labels* (Vlerick Leuven Gent Working Paper Series 2010/07). Ghent, Belgium: Ghent University.

Zampelis, V., Ornstein, E., Franzén, H., & Atroshi, I. (2014). A simple visual analog scale for pain is as responsive as the WOMAC, the SF-36, and the EQ-5D in measuring outcomes of revision hip arthroplasty: A prospective cohort study of 45 patients followed for 2 years. *Acta Orthopaedica*, *85*(2), 128–132.

About the Authors

James R. Lewis
Dr. Lewis is a senior human factors engineer (at IBM since 1981). He has published influential papers in the areas of usability testing and measurement. His books include *Practical Speech User Interface Design* and (with Jeff Sauro) *Quantifying the User Experience*.

Oğuzhan Erdinç
Dr. Erdinç is an associate professor of industrial engineering. He is teaching and studying ergonomics, human computer interaction, usability, and user experience. He has developed and adapted several data collection tools. His research interests include office ergonomics, usability, and user experience.