Int J Speech Technol (2015) 18:479–487
DOI 10.1007/s10772-015-9289-1

Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire

James R. Lewis[1] · **Mary L. Hardzinski**[2]

Received: 30 October 2014 / Accepted: 25 June 2015 / Published online: 5 July 2015

Abstract The Speech User Interface Service Quality (SUISQ) questionnaire is a standardized instrument for the assessment of the usability of interactive voice response (IVR) applications, developed by Polkosky (Toward a social-cognitive psychology of speech technology: affective responses to speech-based e-service, 2005; Mediated interpersonal communication, 2008). During its development, participants rated the quality of recorded interactions rather than interactions in which they participated, leaving open the question of the extent to which the findings would generalize to personal as opposed to observed interactions. The results of a large-scale unmoderated usability study of a natural-language speech recognition IVR demonstrated the utility of the SUISQ for the purpose of assessing personal experiences with service-providing speech user interfaces. The psychometric properties of construct validity and reliability were very similar to those reported by Polkosky. Additional item analyses led to the definition of two subsets of the full set of 25 SUISQ items—a reduced version (SUISQ-R, 14 items) and a maximally-reduced version (SUISQ-MR, 9 items). The SUISQ-R had similar psychometric properties to the full SUISQ, but analysis the SUISQ-MR revealed some weaknesses in its reliability and construct validity. This replication of the original SUISQ findings in a markedly different context of measurement and the availability of a shorter, psychometrically qualified, version of the questionnaire (SUISQ-R) should enhance its utility for usability practitioners who work on the development and assessment of speech-recognition IVRs.

Keywords IVR · Interactive voice response · Subjective assessment of IVR quality · Psychometric evaluation · SUI service quality questionnaire · Usability questionnaire

✉ James R. Lewis
jimlewis@us.ibm.com

Mary L. Hardzinski
mhardzinski@yahoo.com

[1] IBM Corporation, Boca Raton, FL, USA

[2] State Farm Mutual Automobile Insurance Company, Bloomington, IL, USA

1 Introduction

Designers strive to produce usable designs. This is, however, not easy to do—especially in the complex design space of using speech technologies to provide automated service over a telephone (Lewis 2011). One critical aspect of the development of usable speech-enabled interactive voice response (IVR) applications is the measurement of its usability.

The direct measurement of usability is not possible because it is not a property of a person or thing (Lewis 2012; Sauro and Lewis 2012). Usability is an emergent property that depends on the interactions among users, products, tasks and environments, as recognized in the international standard ISO 9241 (International Standards Organization 1998). The ISO standard defines three major components of usability measurement: effectiveness, efficiency, and satisfaction. The first two are performance metrics, typically collected as successful task completions for effectiveness and task completion times for efficiency. Satisfaction, in contrast, is a subjective measurement related to perceived usability, typically collected using a standardized questionnaire.

1.1 Standardization of measurement

A standardized measurement is one for which there is an established procedure for collecting and presenting the

measurement, such as the measurement of time in seconds or temperature in degrees Celsius. Standardized measures have a number of advantages in the practice of science and engineering. Standardized measurements support objectivity of studies and make studies easier to replicate. A number of usability researchers have demonstrated that standardized usability questionnaires are more reliable (more consistently produce the same measurement under the same circumstances) than homegrown or *ad hoc* questionnaires (Hornbæk 2006; Hornbæk and Law 2007; Sauro and Lewis 2009).

The development of standardized measures requires a substantial amount of work. Once developed, however, they are extremely economical. Standardization also makes it easier for practitioners to communicate their results in a way that other practitioners will understand. Standardization also aids the assessment of the generalization of results.

As part of the development of standardized questionnaires, it is the typical practice for the developer to report measurements of its reliability and validity. These are the fundamental elements of psychometric qualification (Nunnally 1978).

1.2 Brief review of psychometric practice

1.2.1 Reliability

Reliability is an assessment of the consistency of a measurement. The most common measurement of a scale's reliability is coefficient alpha (Nunnally 1978), a measure of internal consistency. Coefficient alpha can range from 0 (completely unreliable) to 1 (perfectly reliable). For purposes of research or evaluation in which the final score will be the average of ratings from more than one questionnaire, the typical minimally acceptable reliability is .70 (Landauer 1988; Nunnally 1978).

1.2.2 Validity

Validity refers to the extent to which a measurement actually measures what it claims to measure. There are a number of different approaches to the assessment of validity, including content validity, criterion-related validity, and construct validity. A questionnaire has valid content when the initial pool of items comes from sources that have a rational relationship to the measurement of interest. There are no metrics for content validity; rather, content validity is an outcome of an appropriate process for the creation of candidate items.

Researchers commonly use the correlation coefficient to assess criterion-related validity (the relationship between the measure of interest and a different concurrent or predictive measure). The magnitude of the correlation does not need to be large to provide evidence of validity, but the correlation should be statistically significant. A common minimum criterion for the magnitude of correlations that support the validity hypothesis is .30.

The most common method for assessing construct validity is factor analysis. Factor analysis is a statistical procedure that examines the correlations among variables to discover groups of related variables (Nunnally 1978). Because summated (Likert) scales are more reliable than single-item scores and it is easier to interpret and present a smaller number of scores, it is common to conduct a factor analysis to determine if there is a statistical basis for the formation of measurement scales based on factors. Generally, a factor analysis requires a minimum of five participants per item to ensure stable factor estimates (Nunnally 1978). There are a number of methods for estimating the number of factors in a set of scores when conducting exploratory analyses, including discontinuity and parallel analysis (Cliff 1987; Coovert and McNelis 1988). When previous research has established an expected number of factors, there is a shift of focus from exploratory to confirmatory analysis.

1.2.3 Number of scale steps

Despite the difficulties that individual respondents sometimes have in matching their subjective ratings with scale anchors (Dillman 2000; Sudman et al. 1996; Tourangeau et al. 2000), scale reliability (typically assessed with coefficient alpha) increases as the number of scale steps increases (Nunnally 1978). As the number of scale steps increases from two to twenty, there is an initially rapid increase in reliability that tends to level off at about seven steps (Nunnally 1978). After eleven steps there is very little gain (but no loss) in reliability as the number of steps increases. Lewis (1993) found that mean differences between experimental groups measured with questionnaire items having seven steps correlated more strongly with the observed significance level of statistical tests than did similar measurements using items that had only five scale steps, supporting the use of seven rather than five scale steps.

1.3 Previous research on standardized questionnaires for speech user interfaces

1.3.1 Mean Opinion Scale (MOS)

The MOS questionnaire has been widely used for the assessment of speech heard over a telephone channel and for the assessment of synthetic speech, recommended by the International Telecommunications Union (Schmidt-Nielsen 1995; ITU 1994; van Bezooijen and van Heuven

1997). The MOS is a Likert-style questionnaire, typically with seven 5-point scale items addressing the following TTS characteristics: (1) Global Impression, (2) Listening Effort, (3) Comprehension Problems, (4) Speech Sound Articulation, (5) Pronunciation, (6) Speaking Rate, and (7) Voice Pleasantness. In the most typical use of the MOS, naïve listeners assign scores for each item after listening to speech stimuli, usually sentences (Schmidt-Nielsen 1995). Factor analysis of these items indicated that they supported two underlying constructs: Intelligibility and Naturalness (Kraft and Portele 1995; Lewis 2001).

Polkosky and Lewis (2003) investigated the reliability and validity of the MOS and used psychometric principles to revise and improve the scale. This work resulted in the MOS-Revised (MOS-R). Four subsequent experiments expanded the MOS-R beyond its previous focus on Intelligibility and Naturalness, to include measurement of the Prosody and Social Impression of synthetic voices. The result of this work was the MOS-Expanded (MOS-X), a rating scale shown to be reliable, valid, and sensitive for high-quality evaluation of synthetic speech in applied industrial settings (a total of 15 items, with 4 for Intelligibility, 4 for Naturalness, 3 for Prosody, and 4 for Social Impression). Although the MOS-X and related questionnaires are excellent instruments for their intended purpose, they do not address a sufficient scope for the assessment of the overall perceived usability of an IVR.

1.3.2 Subjective Assessment of Speech System Interfaces (SASSI)

The SASSI (Hone and Graham 2000) is a questionnaire developed for the assessment of users' subjective experiences with speech recognition systems. Starting with an initial pool of 50 items, the final version of the SASSI had 34 items distributed across six scales: System Response Accuracy (9 items), Likeability (9 items), Cognitive Demand (5 items), Annoyance (5 items), Habitability (5 items) and Speed (2 items). The reliabilities of these scales, assessed with coefficient alpha, were respectively .90, .91, .88, .77, .75, and .69. The database of completed SASSI questionnaires used for the psychometric analyses contained 214 questionnaires, collected during usability studies of four different applications.

In contrast to the MOS, the SASSI covers a much broader scope of usability attributes for systems employing speech recognition. A number of researchers have used the SASSI in their evaluations of speech systems. For example, the Association for Computing Machinery (ACM) Digital Library shows over 30 citations from 2005 through 2013. Despite its popularity, there are aspects of the SASSI that reduce its utility when assessing IVR applications. The primary focus of the SASSI is on the consequences of speech input and how it affects perceived usability and affect (positive and negative). Furthermore, a key goal of the SASSI developers was to build a questionnaire that would be generalizable across a broad spectrum of speech applications, from extremely limited use of speech input in small devices to in-car systems to natural language understanding (NLU) queries. Having items applicable across this range of products resulted in a questionnaire that did not address some of the key characteristics of IVR applications.

Due to their common use in enterprise customer service to direct users to skill groups in call centers for human assistance or to automated self-service applications, the assessment of IVRs requires attention to aspects of usability that are not applicable to the broad range of speech-enabled applications. Specifically, the assessment of IVRs requires attention not only to the quality of speech input (the focus of the SASSI) or speech output (the focus of the MOS), but also to the quality of the delivered service, which is one of the key elements of the Speech User Interface Service Quality (SUISQ) questionnaire (Polkosky 2005, 2008).

1.4 The SUISQ questionnaire

The framework for the SUISQ is a questionnaire developed specifically for the assessment of key usability attributes of IVR applications (Polkosky 2005, 2008). An initial pool of 76 items was obtained from the literatures of social psychology, communication, and services marketing. Following several rounds of item and factor analysis, the final version of the SUISQ contained 25 items; factor analysis indicated the presence of four factors, corresponding to its four scales (see Appendix 1). The four scales of the SUISQ (with number of items, estimated reliability, and correlations with customer satisfaction shown in parentheses) are: User Goal Orientation (UGO: 8 items, $\alpha = .92$; $r = .71$), Customer Service Behaviors (CSB: 8 items, $\alpha = .89$, $r = .43$), Speech Characteristics (SC: 5 items, $\alpha = .87$, $r = .40$), and Verbosity (V: 4 items, $\alpha = .69$, $r = -.27$).

The User Goal Orientation items relate to the system's efficiency, user trust, confidence in the system, and clarity of the speech interface. Customer Service Behavior includes items that relate to the friendliness and politeness of the system, its speaking pace, and its use of familiar terms. The Speech Characteristics factor relates to naturalness and enthusiasm of the system voice. Verbosity includes items related to the talkativeness and repetitiveness of the system. In experiments in which participants listened to recordings of interactions with IVR applications and rated them using the SUISQ, Polkosky (2005) found significant correlations between all four metrics and customer satisfaction, with participants preferring higher levels of the first three scales and lower levels of Verbosity.

To obtain the data required to develop the SUISQ, Polkosky (2005) recruited 862 students from the University of South Florida Psychology Department's participant pool (688 females, 161 males, mean age of 20.6), and distributed them about equally among six interface stimuli (Tennis Scoreboard, Directory Dialer, Flights, Movies, Financial Services, and Prescription Refill). Participants listened to a recorded interaction for their assigned stimulus, and then completed a questionnaire that included all the candidate items for the SUISQ plus a variety of concurrent measures including customer satisfaction. To ensure that participants attended to the assigned interaction, they completed a short post-session quiz. The results of the quizzes showed that participants recalled the details of the interactions with reasonable accuracy.

Although she cited precedents in marketing and interpersonal communication studies for using third-party observers to provide ratings of sentiments (e.g., Cargile et al. 1994; Dabholkar and Bagozzi 2002; Patterson 1996), Polkosky stated:

> One of the most important limitations of the present research was the use of observers instead of actual interface users. Findings from social cognition highlight this issue for not only applied speech technology research, but also marketing and interpersonal communication studies, which frequently use observers to generate data on conversational and service interactions. In contrast, findings from the social-cognitive literature warn that interactants and observers may have different affective outcomes. Thus, the present results are limited to observers of speech interface usage and do not necessarily apply to users themselves. This methodological problem has important implications because the use of observers is an efficient and practical means of conducting applied research. … It should be a central goal of future research efforts that potential differences in uses and observer affective responses be explored. (Polkosky 2005, p. 85–86)

Thus, even though findings from the marketing and interpersonal communication literature suggest that observers of interactions might provide ratings of sentiment similar to those who actually experienced the interactions, it is an open research question as to whether participants who actually experience the interaction would provide responses to the SUISQ that would have similar psychometric properties as those reported by Polkosky (2005, 2008).

1.5 Research goals

Our primary research goal was to investigate the psychometric properties of the SUISQ with data collected from participants (callers who actually interacted with the IVR to complete assigned tasks) rather than observers (people who listened to recorded interactions between a caller and an IVR). A secondary research goal was to conduct additional item analyses to explore the reliability and validity of versions of the SUISQ containing fewer than 25 items.

2 Psychometric evaluation of the SUISQ questionnaire

2.1 Method

As part of a larger research effort, 549 employees from a large corporation (415 females, 134 males) volunteered to complete tasks with a test version of a banking IVR using natural-language call routing (Kuo et al. 2003; Lee et al., 2000) using an unmoderated remote usability testing system (Albert et al. 2010). The participants' ages covered a wide span, with 10.6 % from 18 to 29 years old, 24.0 % from 30 to 39, 29.1 % from 40 to 49, 32.1 % from 50 to 59, and 4.2 % over 60. Eighty-five percent of participants used a land-line during the evaluation; 15 % used a cell phone. Over one-third of the participants indicated that they used automated speech systems at least once a week, and about half indicated use once or twice per month. Over half of the respondents indicated that they were *Comfortable* or *Very Comfortable* using these types of systems; just under a third indicated they were *Uncomfortable* or *Very Uncomfortable*.

There were three task groups, each with three different tasks. Participants attempted to complete the tasks in their assigned task group (Group 1, Group 2, or Group 3). The Group 1 tasks were to pay a bill, review transactions from the last three months, and get information about a maturing certificate-of-deposit (CD). For Group 2, the tasks were to update an address, transfer funds, and get information about a health savings account (HSA). The Group 3 tasks were to troubleshoot problems getting into an account, getting the payoff information for a car, and reporting a lost debit card. After completing their assigned group of tasks, participants completed the SUISQ (presented online by the unmoderated usability testing system with items in the order specified by Polkosky, 2008) and provided a rating of satisfaction ("Overall how satisfied are you with your experience using the automated speech system" using a 5-point scale anchored with *Extremely Dissatisfied*, *Dissatisfied*, *Neither Satisfied nor Dissatisfied*, *Satisfied*, and *Extremely Satisfied*). They also indicated via self-report whether they did not accomplish any tasks (Completion = 0), accomplished some tasks (Completion = 1), or accomplished all tasks (Completion = 2).

2.2 Results

These initial analyses investigated the extent to which using the SUISQ as published by Polkosky (2005, 2008), but in this different context of measurement, produced results similar to those in the original research.

2.2.1 Reliability

The estimated reliability for the Overall scale (using all 25 items) was .93. For the specific scales, the values of coefficient alpha were .94 for UGO, .91 for CSB, .78 for SC, and .71 for V. These results were comparable with those reported by Polkosky (2005) which were, respectively, .92 (UGO), .89 (CSB), .87 (SC), and .69 (V) (Polkosky did not report an Overall reliability). The values for coefficient alpha were within .02 for UGO, CSB, and V. The greatest difference was for SC. In both Polkosky (2005) and the present study, however, the reliability exceeded the typical minimum criterion of .70.

2.2.2 Construct validity

Table 1 shows the results of a varimax-rotated principal components analysis of the SUISQ items. Almost all of the items (23 out of 25) aligned with the same component as in Polkosky (2005). Item 7 ("The system was organized and logical") aligned with UGO instead of the expected CSB, and Item 14 ("The system's voice was pleasant") aligned with CSB instead of the expected SC.

2.2.3 Criterion-related validity

Correlations between the SUISQ scales and the satisfaction rating provided evidence of criterion-related (concurrent) validity. Specifically, the correlations (all $p < .01$) were UGO: .74, CSB: .36, SC: .23; and V: −.27. The correlations were within .04 of the values reported by Polkosky (2005) for UGO, CSB, and V. The correlation for SC was lower than that reported by Polkosky (.43), but was still statistically significant and in the same direction.

2.2.4 Sensitivity

As evidence of scale sensitivity, a mixed-model ANOVA using SUISQ Scale as a within-subjects variable and Completion as a between-subjects variable revealed that more successful participants gave significantly better overall SUISQ ratings ($F(2, 519) = 33.7$, $p < .01$), with significant improvement for each higher level of Completion (Bonferroni multiple comparisons, $p < .05$). There was also a significant Scale x Completion interaction ($F(6, 1557) = 24.5$, $p < .01$). As shown in Fig. 1, the effect of Completion on mean rating was strongest for UGO and weakest for SC and V.

3 Psychometric evaluation of the SUISQ-R questionnaire

3.1 Method

Having established a considerable degree of consistency between the psychometric outcomes reported by Polkosky (2005, 2008) and this independent set of data collected in a very different context, it seemed reasonable to analyze the SUISQ items to develop a shorter questionnaire, the SUISQ-Reduced (SUISQ-R). The goal at this stage was to identify 3–4 items per scale that would still have acceptable psychometric properties.

3.2 Results

3.2.1 Item analysis

Analysis of factor loadings and correlations with satisfaction indicated that the items listed in Table 2 should provide sufficient representation for the SUISQ-R to have acceptable psychometric properties (see Appendix 2).

3.2.2 Reliability

The SUISQ-R scales had acceptable reliability as measured with coefficient alpha (UGO: .91, CSB: .88, SC: .80, V: .67, Overall: .88). V had the lowest reliability, under the typical criterion of .70, but just under it. Researchers who require all reliabilities to exceed .70 could modify the SUISQ-R by including all four V items (see Table 1).

3.2.3 Criterion-related validity

The SUISQ-R scales significantly correlated with the rating of satisfaction (UGO: .70, CSB: .32, SC: .21, V: −.32, Overall: .54—all $p < .01$).

3.2.4 Construct validity

A PCA conducted with the items of the SUISQ-R confirmed that the items aligned as expected on the scales.

3.2.5 Sensitivity

For the SUISQ-R scales in the same mixed-model ANOVA as that reported in the previous section, the main effect of Completion ($F(2, 519) = 31.8$, $p < .01$) and the

Table 1 Principal components analysis of the 25 items

Item	Content	UGO	CSB	SC	V
13	I would be likely to use this system again	**.858**	.228	.146	−.124
12	I could trust this system to work correctly	**.834**	.205	.117	−.088
17	I felt confident using this system	**.834**	.245	.159	−.088
10	The system would help me be productive	**.831**	.155	.078	−.089
5	I could find what I needed without any difficulty	**.805**	.190	.031	−.073
3	The system gave me a good feeling about being a customer of this business	**.800**	.180	.162	−.025
1	The system made me feel like I was in control	**.799**	.219	.028	−.098
19	The quality of this system made me want to remain a customer of this business	**.794**	.164	.297	−.105
7*	The system was organized and logical	**.628**	.439	−.009	−.099
6	The system used everyday words	.336	**.758**	.041	−.099
11	The system seemed polite	.256	**.739**	.316	−.105
9	The system spoke at a pace that was easy to follow	.127	**.736**	.079	−.214
14*	The system's voice was pleasant	.188	**.726**	.434	−.084
4	The system used terms I am familiar with	.355	**.711**	−.054	−.041
25	The system seemed professional in its speaking style	.271	**.668**	.400	−.156
23	The system seemed friendly	.290	**.648**	.482	−.150
21	The system seemed courteous	.260	**.599**	.447	−.163
18	The system's voice sounded like a regular person	.096	.139	**.808**	−.054
20	The system's voice sounded natural	.164	.242	**.797**	−.140
24	The system's voice sounded enthusiastic or full of energy	.127	.238	**.658**	−.045
16	The system's voice sounded like people I hear on the radio or television	.027	−.004	**.585**	.121
22	I felt like I had to wait too long for the system to stop talking so I could respond	−.139	−.161	−.036	**.730**
2	The messages were repetitive	−.185	.084	−.084	**.706**
8	The system gave me more details than I needed	.075	−.199	.011	**.701**
15	The system was too talkative	−.223	−.431	.029	**.655**

Factor loadings greater than .500 appear in bold

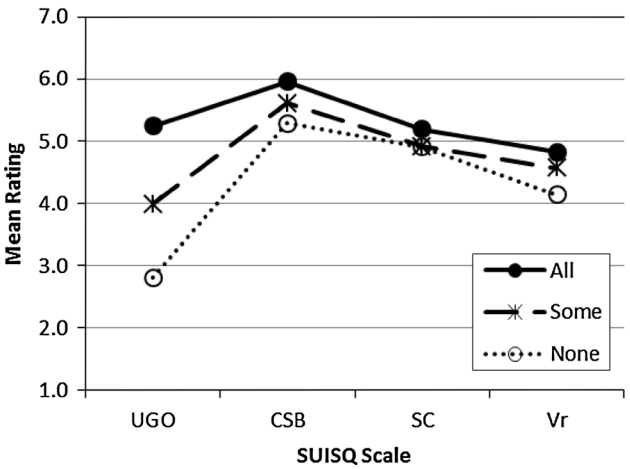

Fig. 1 Scale x Completion interaction. *Note* So all scales have a consistent alignment in the figure with *higher numbers* indicating a better outcome, the scale for V has been reversed using the formula $\underline{V}_r = 8 - \underline{V}$

Completion x Scale interaction ($F(6, 1557) = 18.9$, $p < .0001$) were statistically significant, with patterns of means similar to those in Fig. 1.

4 Psychometric evaluation of the SUISQ-MR questionnaire

4.1 Method

To explore a maximally-reduced version of the SUISQ, we created an initial version that had only two items per scale (the SUISQ-MR).

4.2 Results

4.2.1 Item analysis

Analysis of factor loadings and correlations with satisfaction led to the assignment of the items listed in Table 3 to each scale.

4.2.2 Reliability

The reliabilities of the initial SUISQ-MR scales were estimated as follows using coefficient alpha: UGO: .88, CSB: .75, SC: .68, V: .47.

Table 2 Selected items for the SUISQ-R scales

Scale	Items
UGO	1, 5, 13, 17
CSB	6, 11, 23, 25
SC	18, 20, 24
V	2, 15, 22

Table 3 Preliminary assignment of items to SUISQ-MR scales

Scale	Items
UGO	13, 17
CSB	6, 11
SC	20, 24
V	2, 22

For this maximally-reduced version, two of the scales had estimated reliabilities less than .7. SC was just under the criterion (.02); V was .23 below it. Thus, the final version of the SUISQ-MR included the three V items from the SUISQ-R, with estimated reliability of .67 (see Appendix 3).

4.2.3 *Criterion-related validity*

The scales for the final version of the SUISQ-MR significantly correlated with satisfaction ratings (UGO: .70, CSB: .29, SC: .22, V: −.28, Overall: .55—all $p < .01$).

4.2.4 *Construct validity*

A PCA conducted with the SUISQ-MR items indicated some structural weaknesses. Specifically, the UGO and CSB items coalesced onto one component, and the V items split across two.

4.2.5 *Sensitivity*

The results of the mixed-model ANOVA using the SUISQ-MR scales had outcomes similar to those reported for the SUISQ and SUISQ-R, with a statistically significant main effect of Completion ($F(2, 519) = 32.0$, $p < .0001$) and a significant Completion x Scale interaction ($F(6, 1557) = 19.8$, $p < .0001$).

5 Discussion

The results provided compelling evidence that the SUISQ, developed based on the ratings of observers of speech interactions rather than participants in those interactions, works very well with participants. In addition to this fundamental difference in the measurement context, there were also major differences in backgrounds (students vs. corporate employees), age ranges, and tasks. Despite these differences, the psychometric properties of the SUISQ were strikingly consistent with those originally reported by Polkosky (2005, 2008).

The attempts to develop shorter versions of the SUISQ (SUISQ-R and SUISQ-MR) met with mixed success. The psychometric properties of the 14-item SUISQ-R strongly suggested that it would be an adequate substitution for the full SUISQ. Its only weakness was that the reliability of V was slightly less than .70. It would, however, have rounded up to .7, so this did not seem to be a critical weakness. Researchers who require scales with estimated reliability greater than .70 could accomplish this by using the 4-item version of V from the full SUISQ, resulting in a 15- rather than 14-item instrument.

The 9-item SUISQ-MR, however, suffered from a number of weaknesses. The 2-item version of V had extremely low reliability, which drove the decision to keep the best three items for that scale despite the goal of maximal reduction. The results for criterion-related (concurrent) validity and sensitivity were acceptable. There were some slight weaknesses in the reliability estimates for SC and V. Its greatest problem was its deviation from the expected PCA structure, suggesting a serious weakness in construct validity.

In summary, this replication of the original SUISQ findings in a markedly different context of measurement and the availability of a shorter, psychometrically qualified, version of the questionnaire (SUISQ-R) should enhance its utility for usability practitioners who work on the development and assessment of speech-recognition IVRs. Researchers who need the richest possible set of items for diagnostic purposes should consider using the full SUISQ. Those who absolutely require the shortest possible questionnaire might consider the SUISQ-MR. The SUISQ-R, however, is probably the best choice for most researchers, given its relative brevity and acceptable psychometric properties.

6 Limitations to generalization

One limitation was that our completion metric was based on self-reports rather than an objective assessment of completion due to a limitation of the unmoderated usability test tool in the context of evaluating a phone application. Without downplaying the importance of the perception of task completion, future research would benefit from also obtaining objective measurement of task completion.

Although we had an a priori reason to explore a variety of four-factor solutions based on the results of Polkosky

(2005, 2008), it is possible that the use of one sample for the various versions that we explored may have led to overfitting the model. Also, our data came from an evaluation of one type of voice application, so there is a clear need to replicate the findings using other voice applications, preferably using a variety of input styles (natural language call routing, directed dialog, and touchtone). Successful replication using independent data sets from a variety of voice applications in the future would enhance the generalizability of these findings.

Appendix 1

The Standard SUI service quality (SUISQ) questionnaire

1. The system made me feel like I was in control.
2. The messages were repetitive.
3. The system gave me a good feeling about being a customer of this business.
4. The system used terms I am familiar with.
5. I could find what I needed without any difficulty.
6. The system used everyday words.
7. The system was organized and logical.
8. The system gave me more details than I needed.
9. The system spoke at a pace that was easy to follow.
10. The system would help me be productive.
11. The system seemed polite.
12. I could trust this system to work correctly.
13. I would be likely to use this system again.
14. The system's voice was pleasant.
15. The system was too talkative.
16. The system's voice sounded like people I hear on the radio or television.
17. I felt confident using this system.
18. The system's voice sounded like a regular person.
19. The quality of this system made me want to remain a customer of this business.
20. The system's voice sounded natural.
21. The system seemed courteous.
22. I felt like I had to wait too long for the system to stop talking so I could respond.
23. The system seemed friendly.
24. The system's voice sounded enthusiastic or full of energy.
25. The system seemed professional in its speaking style.

SUISQ scales (based on specification in Polkosky 2005)

User goal orientation (UGO) average items 1, 3, 5, 10, 12, 13, 17, and 19.
Customer service behavior (CSB) average items 4, 6, 7, 9, 11, 21, 23, and 25.
Speech characteristics (SC) average items 14, 16, 18, 20, and 24.
Verbosity (V) average items 2, 8, 15, and 22 (to reverse score: $V_r = 8 - V$).
Overall average of UGO, CSB, SC, and V_r.

Appendix 2

See Table 4.

Table 4 The reduced SUI service quality (SUISQ-R) questionnaire

Item	Original	Scale	Item content
1	13	UGO	I would be likely to use this system again
2	17	UGO	I felt confident using this system
3	5	UGO	I could find what I needed without any difficulty
4	1	UGO	The system made me feel like I was in control
5	6	CSB	The system used everyday words
6	11	CSB	The system seemed polite
7	25	CSB	The system seemed professional in its speaking style
8	23	CSB	The system seemed friendly
9	18	SC	The system's voice sounded like a regular person
10	20	SC	The system's voice sounded natural
11	24	SC	The system's voice sounded enthusiastic or full of energy
12	22	V	I felt like I had to wait too long for the system to stop talking so I could respond
13	2	V	The messages were repetitive
14	15	V	The system was too talkative

SUISQ-R scales

User goal orientation (UGO): average items 1–4

Customer service behavior (CSB): average items 5–8

Speech characteristics (SC): average items 9–11

Verbosity (V): average items 12–14 (to reverse score: $V_r = 8 - V$)

Overall: average of UGO, CSB, SC, and V_r

Appendix 3

See Table 5.

Table 5 The maximally-reduced SUI service quality (SUISQ-MR) questionnaire

Item	Original	Scale	Item content
1	13	UGO	I would be likely to use this system again
2	17	UGO	I felt confident using this system
3	6	CSB	The system used everyday words
4	11	CSB	The system seemed polite
5	20	SC	The system's voice sounded natural
6	24	SC	The system's voice sounded enthusiastic or full of energy
7	22	V	I felt like I had to wait too long for the system to stop talking so I could respond
8	2	V	The messages were repetitive
9	15	V	The system was too talkative

SUISQ-MR scales

User goal orientation (UGO): average items 1–2

Customer service behavior (CSB): average items 3–4

Speech characteristics (SC): average items 5–6

Verbosity (V): average items 7–9 (to reverse score: $V_r = 8 - V$)

Overall: average of UGO, CSB, SC, and V_r

References

Albert, T., Albert, B., & Tedesco, D. (2010). *Beyond the usability lab: Conducting large-scale online user experience studies*. Burlington: Morgan Kaufmann.

Cargile, A., Giles, H., Ryan, E., & Bradac, J. (1994). Language attitudes as a social process: A conceptual model and new directions. *Language & Communication, 14*, 211–236.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego: Harcourt Brace Jovanovich.

Coovert, M. D., & McNelis, K. (1988). Determining the number of common factors in factor analysis: A review and program. *Educational and Psychological Measurement, 48*, 687–693.

Dabholkar, P., & Bagozzi, R. (2002). An attitudinal model of technology-based self-service: Moderating effects of consumer traits and situational factors. *Journal of the Academy of Marketing Science, 30*(3), 184–201.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: John Wiley.

Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering, 6*(3–4), 287–303.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79–102.

Hornbæk, K., & Law, E.L. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of CHI 2007* (pp. 617–626). San Jose: ACM.

International Standards Organization. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability (ISO 9241-11:1998(E))*. Geneva: ISO.

International Telecommunication Union. (1994). *A method for subjective performance assessment of the quality of speech voice output devices (ITU-T recommendation* (p. 85). Geneva: ITU.

Kraft, V., & Portele, T. (1995). Quality evaluation of five German speech synthesis systems. *Acta Acustica, 3*, 351–365.

Kuo, H. J., Siohan, O., & Olive, J. P. (2003). Advances in natural language call routing. *Bell Labs Technical Journal, 7*(4), 155–170.

Landauer, T. K. (1988). Research methods in human-computer interaction. In M. Helander (Ed.), *Handbook of human–computer interaction* (pp. 905–928). New York: Elsevier.

Lee, C.-H., Carpenter, B., Chou, W., Chu-Carroll, J., Reichl, W., Saad, A., & Zhou, Q. (2000). On natural language call routing. *Speech Communication, 31*, 309–320.

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

Lewis, J.R. (2001). Psychometric properties of the mean opinion scale. In *Proceedings of HCI International 2001: Usability Evaluation and Interface Design* (pp. 149–153). Mahwah: Lawrence Erlbaum.

Lewis, J. R. (2011). *Practical speech user interface design*. Boca Raton: Taylor & Francis Group.

Lewis, J. R. (2012). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1267–1312). New York: John Wiley.

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

Patterson, M. L. (1996). Social behavior and social cognition: A parallel process approach. In J. L. Nye & A. M. Brower (Eds.), *What's social about social cognition? Research on socially shared cognition in small groups* (pp. 87–105). Thousand Oaks: Sage.

Polkosky, M. D. (2005). *Toward a social-cognitive psychology of speech technology: Affective responses to speech-based e-service*. Unpublished doctoral dissertation. University of South Florida.

Polkosky, M. D. (2008). Machines as mediators: The challenge of technology for interpersonal communication theory and research. In E. Konjin (Ed.), *Mediated interpersonal communication* (pp. 34–57). New York: Routledge.

Polkosky, M. D., & Lewis, J. R. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology, 6*, 161–182.

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: ACM.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Waltham: Morgan Kaufmann.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett, & S. Greenspan (Eds.), *Applied speech technology* (pp. 195–232). Boca Raton: CRC Press.

Sudman, S., Bradburn, N. M., & Schwartz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass Publishers.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

van Bezooijen, R., & van Heuven, V. (1997). Assessment of synthesis systems. In D. Gibbon, R. Moore, & R. Winski (Eds.), *Handbook of standards and resources for spoken language systems* (pp. 481–563). New York: Mouton de Gruyter.