Voice Interaction Design

The Journal of AVIXD Association for Voice Interaction Design

April 2016 (Vol 1, Issue 1)

Standardized Questionnaires for Voice Interaction Design

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

Abstract

Standardization of measurement is an important aspect of scientific and engineering processes. The benefits of standardization include easier replication, economical reuse, effective communication, and enhanced generalization. The primary purpose of this paper is to provide a summary of the published research on standardized questionnaires suitable for use in the assessment of voice interaction design.

Keywords

voice interaction design, VID, voice user interface design, VUI, speech user interface design, SUI, psychometrics, standardized usability questionnaires, Mean Opinion Scale, MOS, MOS-R, MOS-X, Subjective Assessment of Speech System Interfaces, SASSI, Speech User Interface Service Quality questionnaire, SUISQ, SUISQ-R, SUISQ-MR

Introduction

It is not easy to produce usable designs. This is especially true when developing designs that involve natural technologies such as speech (Lewis, 2011). It is common to think of speech as a natural form of communication, but that very naturalness can contribute to the difficulty of creating usable voice interaction designs. Natural activities, by definition, do not require conscious thought to perform them. Thus, the techniques for designing usable speech user interfaces are not obvious, and therefore benefit from a combination of critically interpreted scientific research and leading design practices (e.g., AVIxD, 2015).

In addition to following leading practice in design, a critical aspect of the development of usable products is the systematic measurement of usability. It is not possible to directly measure usability because it is not a property of a person or thing (Lewis, 2012; Sauro & Lewis, 2012). Rather, it is an emergent property that depends on interactions among users, products, tasks, and environments (ISO, 1998). ISO has defined three major components of usability measurement: effectiveness, efficiency, and satisfaction.

Effectiveness and efficiency are performance metrics. In standard usability testing, the most common effectiveness metric is the successful task completion rate and the most common efficiency metric is successful task completion time. As important as it is to obtain objective performance metrics, it is often also important to assess perceived usability and other subjective characteristics of the systems that people use. This is best accomplished with standardized questionnaires.

Standardized Questionnaires

A questionnaire is a form designed to obtain information from respondents. A common item format for questionnaires is multiple choice, with respondents selecting from a set of alternatives (e.g., "Please select the brand of phone you currently own") or points on a rating scale (e.g., "On a scale of 1 to 5, how satisfied are you with your governor?"). Questionnaires might be designed for a single use or for repeated use in a tracking survey. A standardized questionnaire is one for which there is an established procedure for collecting and presenting the measurement and for which the instrument has undergone psychometric qualification (as described below).

The development of a standardized questionnaire requires substantial effort. After that, however, they are quite economical. Research in usability science (Hornbæk, 2006; Hornbæk & Law, 2007; Sauro & Lewis, 2009) has shown that standardized usability questionnaires have greater reliability than homegrown or ad hoc usability questionnaires. Additional benefits of standardization are (Nunnally, 1978):

- Effective communication among practitioners and researchers
- Enhanced generalizability of findings
- Increased objectivity
- Easier replication

An important aspect of the development of standardized questionnaires is to assess their reliability and validity -- the fundamental elements of psychometric qualification (Nunnallly, 1978).

Brief Review of Psychometric Practice

Reliability

The purpose of reliability analysis is to assess the consistency of a measurement. There are a variety of methods, but the most common for multi-item questionnaires is coefficient alpha, a measure of internal consistency. Coefficient alpha can range from 0 (completely unreliable) to 1 (perfectly reliable). The typical minimum criterion for acceptable reliability for assessments of sentiment (such as standardized usability questionnaires) is 0.70 (Landauer, 1988; Nunnally, 1978).

Validity

For a standardized measurement to be useful, it must not only be reliable, but it must also measure what it claims to measure -- in other words, it must be valid. The assessment of validity takes a number of forms which typically take place at different times during the development of the instrument. At the beginning of development, the method of item selection drives content validity. There is no metric for content validity. Rather, it is assumed as a consequence of starting with an initial pool of items that have rational relationships to the measurement(s) of interest. Those initial items might come from the brainstorming of subject matter experts, from a review of the relevant literature, or both.

Once the questionnaire developers have an initial version of the questionnaire, they begin collecting data -- not only for the questionnaire items but also for other metrics expected to have a relationship with the new metric. Significant correlations between the new metric and the other metrics support claims of concurrent validity.

Construct validity refers to the extent to which the items in a questionnaire group together in the expected pattern. The statistical procedure most often used to assess construct validity is factor analysis. 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 & McNelis, 1988). When previous research (including the research conducted to identify the initial set of items) has established an expected number of factors, there is a shift of focus from exploratory to confirmatory analysis.

Item analysis

Once data has been collected, the next step is to analyze the items to see if it is possible to streamline the questionnaire by deleting the weaker items. One approach to item analysis is to check the alignment of items with the factors computed during factor analysis by examination of the magnitude of the item loadings (similar to correlation coefficients, but with the underlying factors). Items with lower loadings on their factors are candidates for deletion. Another is to examine the correlation between items and related measures hypothesized to measure the same or similar underlying construct or key outcome metrics, keeping items with higher correlations. A third approach is to keep items that discriminate as expected between levels of carefully chosen independent variables. Ideally, these methods would identify the same items as candidates for deletion. If not, then structural considerations (construct validity) generally take priority.

Reassessment of psychometric properties

After eliminating the weaker items, the next step is to collect additional data to ensure that the questionnaire continues to have acceptable levels of reliability and validity. A second round of item analysis can prompt another iteration of the process, but this is not usually necessary.

Development of norms

By itself, a score (individual or average) has no meaning. One way to provide meaning is through comparison, either against a benchmark or via comparison of two sets of data (e.g., different products or different user groups). Another is comparison with norms.

Normative data is collected from one or more representative groups who have completed the questionnaire in a specified setting. Comparison with norms allows assessment of how good or bad a score is, within appropriate limits of generalization. With norms there is always a risk that the new sample doesn't match the normative group(s) (Anastasi, 1976), so it is important to understand where the norms came from when using them to interpret new scores.

Few standardized usability questionnaires have strong normative databases. Those that do typically charge license fees for access to those databases (Sauro & Lewis, 2012). An exception is the System Usability Scale (Brooke, 1996). About ten years after its initial publication several researchers compiled a large database from which it was possible to derive a curved grading scale for mean SUS scores which has proven to be of substantial value to usability practitioners working on graphical, Web, and mobile designs (Sauro & Lewis, 2012).

Summary of standardized questionnaire development

Figure 1 summarizes the development method described above. Think about what you're trying to measure (hypothesize construct(s)), develop a set of items, collect data and assess the items, remove weak items and retest, and, finally, develop norms.

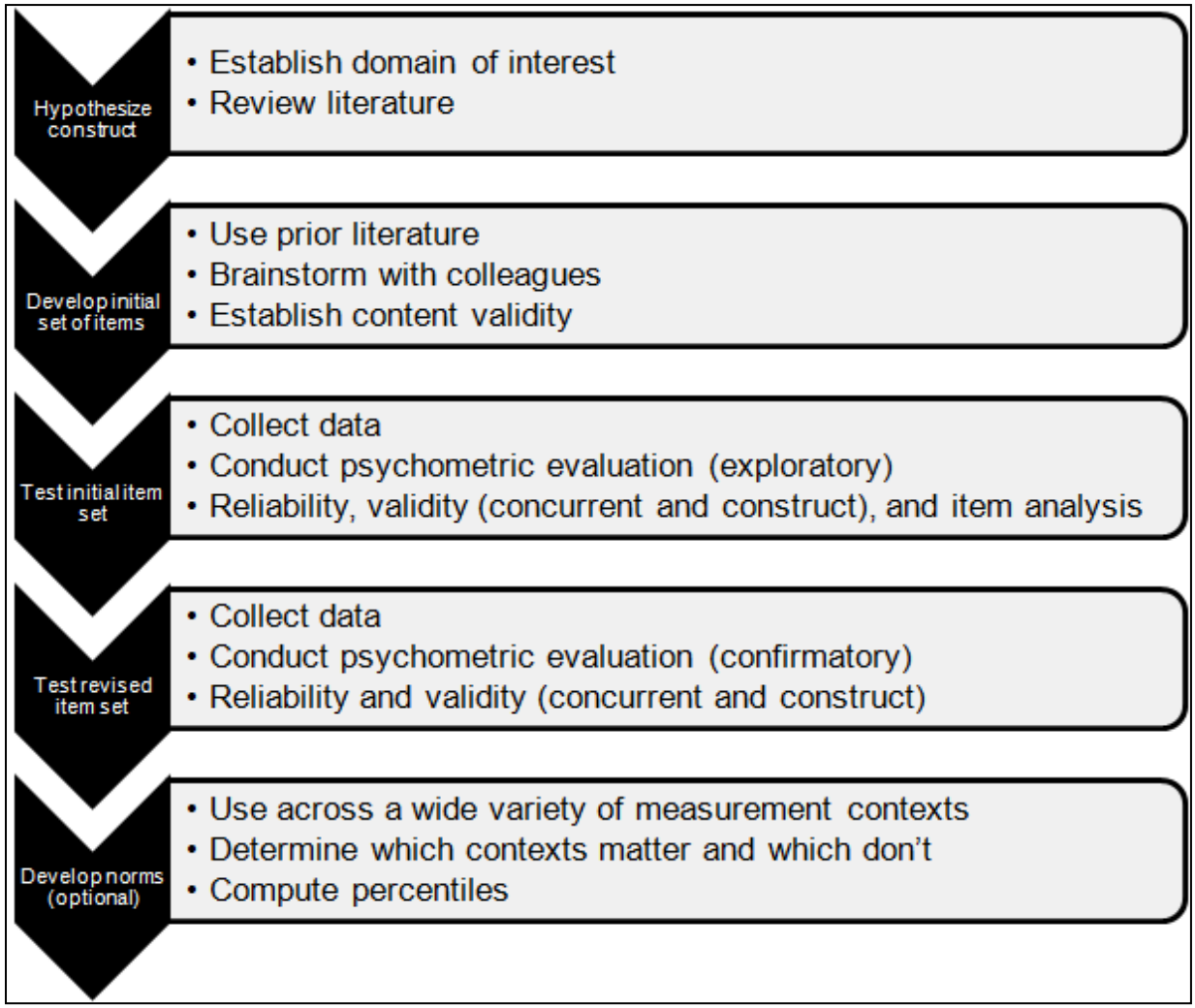

Figure 1. Summary of Standardized Questionnaire Development

Standardized questionnaires for voice interaction design

The focus of the rest of this paper is to describe three standardized questionnaires that are currently available and suitable for use by voice interaction designers:

- The Mean Opinion Scale (MOS)
- Subjective Assessment of Speech System Interfaces (SASSI)
- Speech User Interface Service Quality (SUISQ)

The Mean Opinion Scale (MOS)

Background

Dating from the 1990s, the MOS was originally developed to provide an instrument for the subjective assessment of speech over noisy or otherwise degraded channels, but was adopted for the evaluation of synthetic speech (Schmidt-Nielsen, 1995; ITU, 1994; van Bezooijen and van Heuven, 1997). The most common form of the original version was a questionnaire with seven 5-point items. Although not designed to be a multidimensional metric, factor analysis has typically indicated the two underlying constructs of Intelligibility and Naturalness (Kraft & Portele, 1995; Lewis, 2001). Figure 2 shows the typical MOS items (from Salza, Foti, Nebbia, & Oreglia, 1996) and the factor with which each was associated in Lewis (2001).

Item	Content	1	2	3	4	5	Factor
1	*Global Impression: Your answer must indicate how you rate the sound quality of the voice you have heard.*	Bad	Poor	Good	Fair	Excellent	N
2	*Listening Effort: Your answer must indicate the degree of effort you had to make to understand the message.*	Message not understood with any feasible effort	Major effort required	Effort required	Slight effort required	No effort required	I
3	*Comprehension Problems: Your answer must indicate if you found single words hard to understand.*	Every word	Many	Some	Few	None	I
4	*Speech Sound Articulation: Your answer must indicate if the speech sounds are clearly distinguishable.*	No, not at all	No, not very clear	Fairly clear	Yes, clearly enough	Yes, very clearly	I
5	*Pronunciation: Your answer must indicate if you noticed any anomalies in the naturalness of sentence pronunciation.*	Yes, very annoying	Yes, annoying	Yes, slightly annoying	Yes, but not annoying	No	I
6	*Speaking Rate: Your answer must indicate if you found the speed of delivery of the message appropriate.*	No, too fast	No, too slow	Yes, but faster than preferred	Yes, but slower than preferred	Yes	U
7	*Voice Pleasantness: Your answer must indicate if you found the voice you have heard pleasant.*	Very unpleasant	Unpleasant	Fair	Pleasant	Very pleasant	N

Figure 2. The original Mean Opinion Scale (MOS) -- notations in the Factor column represent Naturalness (N), Intelligibility (I), and Unrelated (U)

Lewis (2001) reported that coefficient alpha for the overall MOS was 0.89, with 0.88 for the Intelligibility factor and 0.81 for the Naturalness factor, all indicative of an acceptable level of reliability. There was also evidence of concurrent validity with paired comparison data and

sensitivity to manipulation (significant differences between ratings for a recorded human voice and two types of text-to-speech voices).

Despite these acceptable psychometric properties, there were a number of weaknesses in the original version of the MOS. Reported validity coefficients were marginally significant, and the failure of the Speaking Rate item to align with any other items could have either been due to its different response item structure or actual independence from the other items. These reported weaknesses spurred additional research.

Polkosky and Lewis (2003) used psychometric principles to revise and improve the MOS with a series of studies that led to the MOS-Expanded (MOS-X), which includes measurement of the prosody and social impression of synthetic voices in addition to their intelligibility and naturalness. The MOS-X has a total of 15 items, with four for Intelligibility, four for Naturalness, three for Prosody, and four for Social Impression (see Figure 3), with 7-point bipolar item formats (7-point items are slightly more reliable than 5-point items -- Lewis, 1993; Nunnally, 1978).

Psychometric Properties of the MOS-X

The evaluation of the final version of the MOS-X (n = 327, between-subjects online assessment of 10 TTS voices) indicated that it had acceptable psychometric properties (Polkosky & Lewis, 2003). Its overall reliability was 0.93, and the coefficient alpha for each factor exceeded 0.85 (Intelligibility: 0.88, Naturalness: 0.86, Prosody: 0.86, and Social Impression: 0.86), demonstrating reliabilities adequate for usability evaluation. Item alignment on factors indicated a high degree of construct validity. MOS-X ratings were sensitive to differences among the 10 TTS voices.

Subjective Assessment of Speech System Interfaces (SASSI)

Background

In 2000, Hone and Graham published the Subjective Assessment of Speech System Interfaces (SASSI) questionnaire. They started with 50 items based on general usability questionnaires, a set of specific speech measures, and a review of the speech system usability literature. Over the course of four separate studies involving a total of eight different speech input systems, they obtained 226 completed questionnaires. Exploratory factor analysis indicated that the items aligned with six factors. After data cleaning and initial item analysis, the published version 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) (see Figure 4). The SASSI has been used in a number of research papers, with over 30 citations from 2005 through 2013 in the Association for Computing Machinery (ACM) Digital Library (e.g., see Hofmann, Ehrlich, Berton, Mahr, Math, & Müller, 2013).

Psychometric Properties

The reliabilities of these scales, assessed with coefficient alpha, were, respectively, 0.90, 0.91, 0.88, 0.77, 0.75, and 0.69. Thus all six scales achieved an acceptable level of reliability (rounding 0.69 to 0.7). There was no information in the initial publication regarding concurrent validity or assessment of sensitivity.

Based on their findings (with fairly small sample sizes), Wechsung, Naumann, and Möller (2008) reported a probable need to revise the SASSI scales. Hone (2014), one of the originators of the SASSI, has also suggested a need for refinement and to more firmly establish its psychometric properties (also see http://people.brunel.ac.uk/~csstksh/sassi.html).

1. *Listening Effort*: Please rate the degree of effort you had to make to understand the message.								
IMPOSSIBLE EVEN WITH MUCH EFFORT	1	2	3	4	5	6	7	NO EFFORT REQUIRED
2. *Comprehension Problems*: Were single words hard to understand?								
ALL WORDS HARD TO UNDERSTAND	1	2	3	4	5	6	7	ALL WORDS EASY TO UNDERSTAND
3. *Speech Sound Articulation*: Were the speech sounds clearly distinguishable?.								
NOT AT ALL CLEAR	1	2	3	4	5	6	7	VERY CLEAR
4. *Precision*: Was the articulation of speech sounds precise?.								
SLURRED OR IMPRECISE	1	2	3	4	5	6	7	PRECISE
5. *Voice Pleasantness*: Was the voice you heard pleasant to listen to?								
VERY UNPLEASANT	1	2	3	4	5	6	7	VERY PLEASANT
6. *Voice Naturalness*: Did the voice sound natural?								
VERY UNNATURAL	1	2	3	4	5	6	7	VERY NATURAL
7. *Humanlike Voice*: To what extent did the voice sound like a human?								
NOTHING LIKE A HUMAN	1	2	3	4	5	6	7	JUST LIKE A HUMAN
8. *Voice Quality*: Did the voice sound harsh, raspy, or strained?								
SIGNIFICANTLY HARSH/RASPY	1	2	3	4	5	6	7	NORMAL QUALITY
9. *Emphasis*: Did emphasis of important words occur?								
INCORRECT EMPHASIS	1	2	3	4	5	6	7	EXCELLENT USE OF EMPHASIS
10. *Rhythm*: Did the rhythm of the speech sound natural?								
UNNATURAL OR MECHANICAL	1	2	3	4	5	6	7	NATURAL RHYTHM
11. *Intonation*: Did the intonation pattern of sentences sound smooth and natural?								
ABRUPT OR ABNORMAL	1	2	3	4	5	6	7	SMOOTH OR NORMAL
12. *Trust*: Did the voice appear to be trustworthy?								
NOT AT ALL TRUSTWORTHY	1	2	3	4	5	6	7	VERY TRUSTWORTHY
13. *Confidence*: Did the voice suggest a confident speaker?								
NOT AT ALL CONFIDENT	1	2	3	4	5	6	7	VERY CONFIDENT
14. *Enthusiasm*: Did the voice seem to be enthusiastic?								
NOT AT ALL ENTHUSIASTIC	1	2	3	4	5	6	7	VERY ENTHUSIASTIC
15. *Persuasiveness*: Was the voice persuasive?								
NOT AT ALL PERSUASIVE	1	2	3	4	5	6	7	VERY PERSUASIVE

Figure 3. The Mean Opinion Scale-Expanded (MOS-X) -- the four factors are Intelligibility (Items 1-4), Naturalness (Items 5-8), Prosody (Items 9-11), and Social Impression (Items 12-15) -- factor scores are the means of their item scores; the overall score is the mean of the factor scores

The SASSI

	Item	Strongly disagree	Disagree	Slightly disagree	Neutral	Slightly agree	Agree	Strongly agree
System Response Accuracy	1. The system is accurate.	O	O	O	O	O	O	O
	2. The system is unreliable.	O	O	O	O	O	O	O
	3. The interaction with the system is unpredictable.	O	O	O	O	O	O	O
	4. The system didn't always do what I wanted.	O	O	O	O	O	O	O
	5. The system didn't always do what I expected.	O	O	O	O	O	O	O
	6. The system is dependable.	O	O	O	O	O	O	O
	7. The system makes few errors.	O	O	O	O	O	O	O
	8. The interaction with the system is consistent.	O	O	O	O	O	O	O
	9. The interaction with the system is efficient.	O	O	O	O	O	O	O
Likeability	10. The system is useful.	O	O	O	O	O	O	O
	11. The system is pleasant.	O	O	O	O	O	O	O
	12. The system is friendly.	O	O	O	O	O	O	O
	13. I was able to recover easily from errors.	O	O	O	O	O	O	O
	14. I enjoyed using the system.	O	O	O	O	O	O	O
	15. It is clear how to speak to the system.	O	O	O	O	O	O	O
	16. It is easy to learn to use the system.	O	O	O	O	O	O	O
	17. I would use this system.	O	O	O	O	O	O	O
	18. I felt in control of the interaction with the system.	O	O	O	O	O	O	O
Cognitive Demand	19. I felt confident using the system.	O	O	O	O	O	O	O
	20. I felt tense using the system.	O	O	O	O	O	O	O
	21. I felt calm using the system.	O	O	O	O	O	O	O
	22. A high level of concentration is required when using the system.	O	O	O	O	O	O	O
	23. The system is easy to use.	O	O	O	O	O	O	O
Annoyance	24. The interaction with the system is repetitive.	O	O	O	O	O	O	O
	25. The interaction with the system is boring.	O	O	O	O	O	O	O
	26. The interaction with the system is irritating.	O	O	O	O	O	O	O
	27. The interaction with the system is frustrating.	O	O	O	O	O	O	O
	28. The system is too inflexible.	O	O	O	O	O	O	O
Habitability	29. I sometimes wondered if I was using the right word.	O	O	O	O	O	O	O
	30. I always knew what to say to the system.	O	O	O	O	O	O	O
	31. I was not always sure what the system was doing.	O	O	O	O	O	O	O
	32. It is easy to lose track of where you are in an interaction with the system.	O	O	O	O	O	O	O
Speed	33. The interaction with the system is fast.	O	O	O	O	O	O	O
	34. The system responds too slowly.	O	O	O	O	O	O	O

Figure 4. The Subjective Assessment of Speech System Interfaces (SASSI) questionnaire

Speech User Interface Service Quality (SUISQ)

Background

The SUISQ is a questionnaire developed for the assessment of important usability attributes of Interactive Voice Response (IVR) applications (Polkosky, 2005, 2008). Because IVRs are typically part of an enterprise's customer service offerings, one of the unique aspects of the SUISQ is its inclusion of items related to satisfactory customer service. Polkosky obtained an initial pool of 76 items from the literatures of social psychology, communication, and services marketing, using iterations of factor analysis and item analysis to arrive at the final version with 25 items aligning with four factors: User Goal Orientation (UGO: 8 items), Customer Service Behaviors (CSB: 8 items), Speech Characteristics (SC: 5 items), and Verbosity (V: 4 items). Figure 5 depicts the original version of the SUISQ (25 agreement items using 7-point scales).

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

Figure 5. The original version of the SUISQ -- the four factors are User Goal Orientation (Items 1, 3, 5, 10, 12, 13, 17, 19), Customer Service Behavior (Items 4, 6, 7, 9, 11, 21, 23, 25), Speech Characteristics (Items 14, 16, 18, 20, 24), and Verbosity (Items 2, 8, 15, 22) -- factor scores are the means of their item scores; the overall score is the mean of the factor scores after reversing the Verbosity mean using the formula Vr = 8 – V

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.

Polkosky (2005) obtained her data by having participants (862 college students) listen to recordings of users interacting with one of six speech systems and then having the participants rate those interactions. The marketing and interpersonal communication literatures suggest that observers of interactions provide ratings of sentiment similar to those who actually experienced the interactions (Cargile, Giles, Ryan, & Bradac, 1994; Dabholkar & Bagozzi, 2002; Patterson, 1996). At the time, however, it was an open research question as to whether

participants who actually experienced the interaction would provide SUISQ ratings with similar psychometric properties.

Ten years later, Lewis and Hardzinski (2015) published their use of the standard version of the SUISQ in a large-scale (*n* = 549 employees of a large corporation) unmoderated usability study (Albert, Tullis, & Tedesco, 2010) of a natural-language speech recognition IVR. Participants completed tasks with a test version of a banking IVR that used natural-language call routing (Kuo, Siohan, & Olive, 2003; Lee et al., 2000). 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 and provided a rating of satisfaction. 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).

The psychometric properties of reliability, concurrent validity, and construct validity were very similar to those reported by Polkosky (2005), with 23 of 25 items aligning as expected. Lewis and Hardzinski (2015) also conducted additional item analyses to further streamline the questionnaire, publishing the Reduced SUI Service Quality (SUISQ-R) and Maximally Reduced SUI Service Quality (SUISQ-MR) questionnaires (see Figures 6 and 7). Because the reliability of Verbosity was very poor with only the two best items, the SUISQ-MR has the same three-item Verbosity scale as SUISQ-R.

	The SUISQ-R	Strongly Disagree							Strongly Agree
			1	2	3	4	5	6	7
1	I would be likely to use this system again.		O	O	O	O	O	O	O
2	I felt confident using this system.		O	O	O	O	O	O	O
3	I could find what I needed without any difficulty.		O	O	O	O	O	O	O
4	The system made me feel like I was in control.		O	O	O	O	O	O	O
5	The system used everyday words.		O	O	O	O	O	O	O
6	The system seemed polite.		O	O	O	O	O	O	O
7	The system seemed professional in its speaking style.		O	O	O	O	O	O	O
8	The system seemed friendly.		O	O	O	O	O	O	O
9	The system's voice sounded like a regular person.		O	O	O	O	O	O	O
10	The system's voice sounded natural.		O	O	O	O	O	O	O
11	The system's voice sounded enthusiastic or full of energy.		O	O	O	O	O	O	O
12	I felt like I had to wait too long for the system to stop talking so I could respond.		O	O	O	O	O	O	O
13	The messages were repetitive.		O	O	O	O	O	O	O
14	The system was too talkative.		O	O	O	O	O	O	O

Figure 6. The reduced version of the SUISQ -- the four factors are User Goal Orientation (Items 1-4), Customer Service Behavior (Items 5-8), Speech Characteristics (Items 9-11), and Verbosity (Items 12-14) -- factor scores are the means of their item scores; the overall score is the mean of the factor scores after reversing the Verbosity mean using the formula Vr = 8 – V

	The SUISQ-MR	Strongly Disagree						Strongly Agree
		1	2	3	4	5	6	7
1	I would be likely to use this system again.	O	O	O	O	O	O	O
2	I felt confident using this system.	O	O	O	O	O	O	O
3	The system used everyday words.	O	O	O	O	O	O	O
4	The system seemed polite.	O	O	O	O	O	O	O
5	The system's voice sounded natural.	O	O	O	O	O	O	O
6	The system's voice sounded enthusiastic or full of energy.	O	O	O	O	O	O	O
7	I felt like I had to wait too long for the system to stop talking so I could respond.	O	O	O	O	O	O	O
8	The messages were repetitive.	O	O	O	O	O	O	O
9	The system was too talkative.	O	O	O	O	O	O	O

Figure 7. The maximally reduced version of the SUISQ -- the four factors are User Goal Orientation (Items 1-2), Customer Service Behavior (Items 3-4), Speech Characteristics (Items 5-6), and Verbosity (Items 7-9) -- factor scores are the means of their item scores; the overall score is the mean of the factor scores after reversing the Verbosity mean using the formula Vr = 8 – V

Psychometric Properties

Table 1 shows the reliabilities (coefficient alpha) and Table 2 the concurrent validities (correlation with a rating of general satisfaction) from the various SUISQ analyses. Reliabilities reported by Polkosky (2005) and Lewis and Hardzinski (2015) for the original questionnaire were almost identical (within .02 for each scale). Scale reliabilities were generally adequate, even for the maximally reduced version. Values of coefficient alpha for Verbosity were consistently at or just below .70 (marginally acceptable).

Table 1. SUISQ Reliability Analyses

Reliability (Coefficient Alpha)	UGO	CSB	SC	V	Overall
SUISQ Original (Polkosky, 2005)	0.92	0.89	0.87	0.69	na
SUISQ Original (Lewis & Hardzinski, 2015)	0.94	0.91	0.87	0.71	0.93
SUISQ-R (Lewis & Hardzinski, 2015)	0.91	0.88	0.80	0.67	0.88
SUISQ-MR (Lewis & Hardzinski, 2015)	0.88	0.75	0.68	0.67	0.80

Note: UGO = User Goal Orientation, CSB = Customer Service Behavior, SC = Speech Characteristics, V = Verbosity

Table 2. SUISQ Concurrent Validity Analyses (correlations with satisfaction)

Concurrent Validity	UGO	CSB	SC	V	Overall
SUISQ Original (Polkosky, 2005)	0.71	0.43	0.40	-0.27	na
SUISQ Original (Lewis & Hardzinski, 2015)	0.74	0.36	0.23	-0.27	0.57
SUISQ-R (Lewis & Hardzinski, 2015)	0.70	0.32	0.21	-0.32	0.54
SUISQ-MR (Lewis & Hardzinski, 2015)	0.70	0.29	0.22	-0.32	0.55

Note: UGO = User Goal Orientation, CSB = Customer Service Behavior, SC = Speech Characteristics, V = Verbosity

All correlations in Table 2 were statistically significant ($p < .01$). For the original versions of the SUISQ, the correlations from the Polkosky (2005) and Lewis and Hardzinski (2015) analyses were very consistent for UGO, CSB, and V; the correlations for SC were of different magnitude (but in the same positive direction). Across the three versions examined by Lewis and Hardzinski (2015), there were similar magnitudes of correlations for the various scales.

Despite the differences in data collection protocols, the factor analyses of the original version of the SUISQ from Polkosky (2005) and Lewis and Hardzinski (2015) were similar (with 23 of 25 items aligning on the same factors), providing evidence of construct validity. Sensitivity analysis showed that more successful participants rated their interaction more highly, with similar outcomes for each version of the questionnaire.

Discussion

This paper has reviewed the intended use and psychometric properties of three standardized questionnaires that have potential application in the assessment of voice user interfaces: the MOS, SASSI, and SUISQ. There is currently only one version of the SASSI, but there are several versions of the MOS and SUISQ. Although there may be legitimate reasons to use alternate versions, considering tradeoffs between breadth and brevity, the preferred versions are the MOS-X and SUISQ-R. Table 3 summarizes the psychometric properties of these three questionnaires.

Table 3. Summary of Psychometric Properties of the Three Standardized Questionnaires

Characteristic	**MOS-X**	**SASSI**	**SUISQ-R**
Designed to assess	Voice quality	General speech system usability	IVR usability
Number of factors	4	6	4
Overall reliability	0.93	NA	0.88
Subscales	Intelligibility, Naturalness, Prosody, Social Impression	System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability, Speed	User Goal Orientation, Customer Service Behavior, Speech Characteristics, Verbosity
Subscale reliabilities (respective)	0.88, 0.86, 0.86, 0.86	0.90, 0.91, 0.88, 0.77, 0.75, 0.69	0.91, 0.88, 0.80, 0.67
Construct validity	Yes	Yes	Yes
Concurrent validity	NA	NA	Yes
Evidence of sensitivity	Yes	NA	Yes
Availability of norms	No	No	No

The available data generally support the use of the questionnaires for their intended purposes, but inspection of Table 3 shows some gaps in their psychometric qualification. All are multifactor instruments with published subscale reliabilities and evidence of construct validity. There is no published method for computing an overall score for the SASSI and no assessment

of its overall reliability. Neither the MOS-X nor the SASSI have published evidence of concurrent validity. Finally, there are no published norms for any of the questionnaires.

As mentioned in the Introduction, the absence of published norms does not render the questionnaires useless, but the availability of norms would dramatically increase their value. Assuming the use of the questionnaires in future research, it will be relatively easy to fill many of the gaps, but the development of compelling norms is much more difficult.

The best known example of the emergence of norms for a standardized usability questionnaire is for those that have recently appeared in print for the System Usability Scale (SUS). Note that the SUS was developed at DEC in the mid 1980s. Roughly 10 years later it was published (Brooke, 1996), and 16 years after that saw the publication of norms based on data from almost 450 studies (Sauro & Lewis, 2012) -- a process that took decades, and without an extraordinary effort, would still not be available.

The norms published by Sauro and Lewis (2012) came primarily from data accumulated by Jeff Sauro over a decade. In addition to his own use of the SUS, he developed contacts throughout the usability community and collected many anonymized SUS datasets from industrial usability studies. As the new journal Voice Interaction Design launches, it is intriguing to speculate that it could become the venue in which researchers and practitioners might publish data collected using standardized questionnaires such as those reviewed in this paper which, over time, could lead to the development of compelling norms for the interpretation of their scores.

Recommendations

As a discipline, we have a good start on the development of standardized questionnaires suitable for use in the assessment of voice user interfaces, but there is still work to do.

- As appropriate for their purposes, practitioners and researchers should use existing standardized questionnaires.
- Unless there is a compelling reason to do otherwise, strongly consider using the MOS-X and SUISQ-R versions of these questionnaires.
- When using the SASSI, consider sharing the results with its developers (for more information, see http://people.brunel.ac.uk/~csstksh/sassi.html).
- When possible, practitioners and researchers should publish the results of using the questionnaires, including publication of the full set of cases (in appendices, anonymized if necessary) to support future psychometric analyses, including the eventual development of norms.
- In addition to using the standardized questionnaires provided in this paper, researchers should consider also collecting SUS data and other standard outcome metrics (e.g., overall experience, likelihood to recommend) to improve our understanding of the relationships between the MOS, SASSI, and SUISQ and these other metrics (and to provide consistent opportunities to compute concurrent correlations).

Tips for Voice Interaction Designers

Regarding the use of standardized questionnaires in voice interaction design:

- Voice interaction designers should be aware of the existing standardized questionnaires, their intended uses, and their psychometric qualities.
- When planning evaluations of voice user interfaces, designers should ensure that the team members planning the evaluations know about the questionnaires and of the value of using them.

Conclusion

Standardization of measurement is an important aspect of scientific and engineering processes. It takes a substantial amount of work to develop standardized questionnaires, but once developed, they are easy to reuse. The primary purpose of this paper was to provide a summary of the published research on standardized questionnaires suitable for use in the assessment of voice interaction design -- specifically, the MOS, SASSI, and SUISQ. To as great

an extent as possible, practitioners and researchers should use these standardized questionnaires when assessing applications that use voice user interfaces and should publish their results to support continuing evaluation of their psychometric properties in various contexts and the potential eventual development of norms.

References

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

Anastasi, A. (1976). *Psychological testing*. New York, NY: Macmillan.

AVIxD (2015). *Voice interaction design wiki*. Available online at <http://videsign.wikispaces.com/>.

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.

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

Cliff, N. (1987). *Analyzing multivariate data*. San Diego, CA: 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.

Hofmann, H., Ehrlich, U., Berton, A., Mahr, A., Math, R., & Müller, C. (2013). Evaluation of speech dialog strategies for Internet applications in the car. In *Proceedings of the SIGDIAL 2013 Conference* (pp. 233–241). Metz, France: Association for Computational Linguistics.

Hone, K. S. (2014). Usability measurement for speech systems: SASSI revisited. In *Proceedings of CHI 2014: Designing Speech and Language Interactions Workshop* (pp. 1-4). Toronto, Canada: ACM.

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, CA: 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, Switzerland: 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, Switzerland: 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, NJ: Lawrence Erlbaum.

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

Lewis, J.R. (2012). Usability testing. In Salvendy, G. (Ed.), *Handbook of Human Factors and Ergonomics* (pp. 1267-1312). New York, NY: John Wiley.

Lewis, J. R., & Hardzinski, M. L. (2015). Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology*, (in press, available online as DOI 10.1007/s10772-015-9289-1).

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, CA: 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, NY: 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.

Salza, P.L., Foti, E., Nebbia, L., and Oreglia, M. (1996). MOS and pair comparison combined methods for quality evaluation of text to speech systems. *Acta Acustica*, *82*, 650–656.

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. (2012). *Quantifying the user experience: Practical statistics for user research*. Burlington, MA: Morgan Kaufmann.

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

Weiss, B., Wechsung, I., Naumann, A., & Möller, S. (2008). Subjective evaluation method for speech-based uni- and multimodal applications. *Lecture Notes in Computer Science*, *5078*, 285-288.

van Bezooijen, R. & van Heuven, V. (1997). Assessment of synthesis systems. In D. Gibbon, R. Moore, and R. Winski (Eds.), *Handbook of Standards and Resources for Spoken Language Systems* (pp. 481-563). New York, NY: Mouton de Gruyter.

About the Author

James R. (Jim) Lewis
Jim is human factors engineer at IBM, specializing in voice interaction design and usability assessment. He is a past president of AVIxD. His books include *Practical Speech User Interface Design* (2011) and (with Jeff Sauro in 2012), *Quantifying the User Experience*.