Design and Psychometric Evaluation of an American Sign Language Translation of the System Usability Scale

Matt Huenerfauth
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14618 USA
matt.huenerfauth@rit.edu

Kasmira Patel
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14618 USA
kxp3157@rit.edu

Larwan Berke
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14618 USA
larwan.berke@mail.rit.edu

ABSTRACT
In usability studies, designers and researchers frequently use subjective questions to evaluate participants' impression of the usability of some product. The System Usability Scale (SUS) is a popular standardized questionnaire consisting of ten English statements about the usability of a product, to which participants indicate their agreement on a five-point scale. Many deaf adults in the U.S. have lower levels of English reading literacy, but there are currently no standardized questionnaires similar to SUS for Deaf and Hard-of-Hearing (DHH) users who are fluent in American Sign Language (ASL). To facilitate the inclusion of such users in studies, we created an ASL translation of SUS following accepted methods of survey translation: using a bilingual team including native ASL signers who are members of the Deaf community, along with back-translation evaluation to determine whether the meaning of the original was preserved. To validate whether key psychometric properties were preserved during translation, we deployed the ASL instrument in a study with 30 DHH participants. By comparing the results to users' responses to another measurement instrument, along with scores from 10 additional DHH participants responding to the original English SUS, we verified the criterion validity and internal reliability of the new "ASL-SUS." We are disseminating the translated instrument to promote the inclusion of DHH users in HCI research studies or in usability testing of consumer products.

CCS Concepts
• Human-centered computing~Empirical studies in accessibility • Human-centered computing~Accessibility design and evaluation methods

Keywords
System Usability Scale, SUS, American Sign Language, Translation, ASL-SUS, Criterion Validity, Internal Reliability

1. INTRODUCTION
To ensure that technology is accessible to diverse users, researchers and designers should ideally include people with disabilities during evaluation studies: The focus of our research is on finding ways to make it easier for researchers to include people who are Deaf or Hard-of-Hearing (DHH) in such studies. Often, designers will gather feedback from study participants by asking them to respond to questions about their impression of a system, and many studies will use pre-existing questionnaires for this purpose.

One popular instrument is the System Usability Scale (SUS), which consists of a ten-item Likert scale with English statements about the usability of some product [5]. For each item, participants indicate their agreement on a five-point scale from Strongly Disagree to Strongly Agree. From these responses, researchers use a rubric to calculate a final score on a range from 0 to 100, with higher scores indicating that the participant believed the product to be very usable. SUS has been used to evaluate a wide range of products, including hardware, computer software, websites, and mobile applications. Given the ease with which it is administered and scored, it has become a ubiquitous measurement instrument in HCI research and in commercial HCI usability testing.

One logistical challenge that designers face when including DHH users in empirical studies is that many people who are DHH prefer to communicate using American Sign Language (ASL). In fact, studies indicate that there are a half-million people who consider ASL as a primary means of communication [23]. Previous research on English literacy among U.S. deaf adults has found that many have lower literacy than their hearing peers [29]. In a usability study, DHH participants with lower English literacy may not fully understand questions presented in the form of English text, which may lead to discomfort or responses which are less reliable [13].

We therefore investigate how to translate SUS, one of the most commonly used usability questionnaires, into ASL. By making this translated instrument (in the form of ASL videos with translated instructions and question items) available to the HCI research community, our goal is to enable researchers to more easily include DHH participants in their studies, alongside hearing participants. In addition to creating this ASL translation, we have also conducted an evaluation to determine whether it preserves the meaning and other characteristics of the original English version, to enable researchers to compare scores across both versions.

This paper is organized as follows: Section 2 surveys prior work on HCI research with DHH users, standardized questionnaires, and prior efforts to translate them into other languages. Sections 3 and 4 outline our research questions and methodology for creating and evaluating the ASL version. Section 5 presents our evaluation results, and Section 6 summarizes our conclusions and future work.

2. LITERATURE SURVEY

2.1 Technology Research with DHH Users
It is difficult to quantify how often people with disabilities are *excluded* from usability studies or research due to the perception among some researchers that logistical barriers to inclusion are too

ASSETS'17, October 29-November 1, 2017, Baltimore, MD, USA

ACM ISBN 978-1-4503-4926-0/17/10…$15.00
https://doi.org/10.1145/3132525.3132540

great, especially when the study's focus is on the general population rather than on people with disabilities specifically. As discussed in [27], while HCI researchers generally agree that the participants in a study should be representative of the population of users, there can be a variety of challenges in recruiting participants with disabilities. People with disabilities are often not included in user testing, which can lead to inaccurate results [27]. In fact, a popular research-methods textbook for training future HCI professionals devotes an entire chapter to motivating students that people with disabilities should be included in studies and explaining how to overcome logistical barriers to doing so [19]. Some researchers include proxy users, e.g. sighted users wearing blindfolds or hearing users with the computer muted, rather than recruit users with disabilities [19, 27]. Others have examined the use of remote testing for people with disabilities, which addresses some barriers to recruitment due to population sparsity or difficulty traveling, but remote testing but does not address the English literacy issues of many DHH users (discussed above), unless steps are taken to provide question items using ASL videos, e.g. as in [26, 28].

Computing accessibility researchers publishing their work at competitive research conferences or journals will generally include people with disabilities in their empirical studies [27]. For example, researchers studying technologies related to sign language animation or recognition often conduct evaluation studies with participants who are DHH [10, 16]. Technology researchers who are studying current challenges or barriers faced by people who are DHH will also include DHH participants in studies: For instance, researchers have examined the communication challenges of DHH users in business or educational contexts [8, 11, 17].

Research on DHH users has tended to employ non-standardized survey questions, created specifically for each particular project; often, these question items are provided in the form of ASL videos, sometimes alongside English text versions of the items. For instance, one online survey of DHH students included both English and ASL versions of each question, to study the communication approaches used during meetings with hearing peers [11]. Another online survey used ASL videos to present question items, in order to collect judgments from ASL signers about the intelligibility of ASL videos of various levels of quality [28].

Some prior accessibility researchers have created and disseminated standardized survey instruments in the form of ASL videos. In [16], researchers sought to identify relationships between DHH participants' opinions about ASL animation technology and their responses to standardized surveys about technology use (originally designed for older users). In support of this research, the authors created ASL translations of pre-existing English survey questions about technology use and displayed these items to their participants in the form of ASL videos. However, the authors did not conduct a formal assessment of the quality of their ASL translations. In [15, 16], researchers disseminated sets of question items (in the form of both English text and ASL video) that could be used to evaluate the quality of ASL animation generation systems, to promote a standardized set of evaluation instruments among that community to enable comparisons of results across studies. However, none of this prior research produced translations of commonly used HCI usability survey instruments, nor did these prior studies include a formal assessment of translation quality.

A key premise of our work is that if there were greater availability of question items (available as both English text and ASL videos) suitable for evaluating usability of software or websites, more HCI researchers may include of DHH participants in their studies, even when the study is not specifically focused on DHH users.

2.2 Standardized Usability Questionnaires

When conducting a study to measure participants' subjective opinion about the usability of some technology, HCI researchers often use *standardized* questionnaires, i.e. official, accepted versions of questionnaires with a commonly agreed-upon set of questions, presented in a particular order, and with rubrics for how to calculate a final score based on a participant's responses. When the same questionnaire is used, it is easier to compare results across studies. Further, it is possible to archive participant scores across a large number of studies, perhaps subcategorized as to the type of technology evaluated. This set of published response values (or tables containing mean and standard deviations) is referred to as "norms," and the existence of this large quantity of historical data makes it easier to evaluate the results of a new study. That is, after a researcher has used a standardized instrument to produce a score representing the usability of some technology, they can compare their participants' responses to these norms to understand how their new results compare, in general, to prior studies. Often, the items included in a standardized questionnaire were chosen from an initially larger pool of questions, through a process of psychometric validation, to create a measurement instrument with a set of desirable psychometric properties [20, 25]. Since we discuss some properties in this paper, we provide brief definitions below for how we use some terms; detailed definitions appear in [1, 20, 25].

- **Construct Validity** refers to whether the instrument actually measures the real-world phenomena it is meant to, e.g. usability; this property is often assessed by examining the internal factor structure, i.e. whether subsets of questions on the questionnaire are correlated and whether these clusters relates to specific sub-scores or factors to be measured.
- **Criterion Validity** refers to whether the instrument relates to some other external property that can be measured; this property is often assessed based on whether the resulting score is correlated with some other trusted measure that respondents complete concurrently (i.e. another instrument they complete at the same time) or to its ability to predict some outcome.
- **Content Validity** refers to whether the questionnaire covers all of the essential aspects of the real-world phenomena it is supposed to measure; this property is often assessed by asking a panel of expert judges to evaluate the individual items.
- **Internal Reliability** refers to its internal consistency, i.e. whether all of the items in the instrument contribute equally to its score; this property is often assessed by calculating the Cronbach's alpha coefficient for the group of items [7].

2.3 Translating Surveys to Other Languages

It is known that asking participants to respond to a questionnaire in their non-native language affects their response scores [13]. So, researchers have established customary methods for translating a standardized survey into another language and evaluating whether the meaning of the original was preserved [13]. Further, researchers who translate a standardized questionnaire often evaluate the new version to determine whether certain desirable psychometric properties of the original (section 2.2) were preserved during the translation process. This section describes prior work in translating standardized usability questionnaires to other languages and the methods used to evaluate the success of those efforts. The four projects summarized below are the primary focus of this section:

- **English-to-Persian SUS:** Researchers in Iran translated the SUS into Persian, with a goal of creating an Iranian version that retained the psychometric properties of the original [9]. To ensure that the content of the new instrument was valid, a

panel of 10 experts first analyzed the Persian version of the questions. Next, the authors conducted a study with 202 university students to evaluate the resulting survey scores.

- **English-to-Slovene SUS:** Researchers translated SUS into the Slovene language [4]. To evaluate the translation, the authors assembled a committee of reviewers with multidisciplinary backgrounds. The authors also evaluated their new instrument in a user study with 182 participants.
- **English-to-Turkish CSUQ:** Researchers translated the Computer System Usability Questionnaire (CSUQ) from English into Turkish [12]. While the focus of our project is SUS, CSUQ is a similar standardized usability questionnaire; so, the process by which researchers translated and evaluated it is relevant for us to consider. After conducting their translation, the authors evaluated the validity of the Turkish version by using the instrument in a user study with 97 participants to evaluate its psychometric properties.
- **English-to-German SUS:** Researchers authored a German translation of SUS based, in part, on an online crowdsourcing effort by a large group of volunteers [22]. However, no formal evaluation was conducted of the quality of the resulting translation nor of its psychometric properties.

We analyze these prior efforts to translate standardized usability surveys into other languages along several dimensions, which correspond to the three sub-sections below. Throughout this survey, key methodological elements of these prior studies are highlighted in **bold** font, and the research methodology for our translation of SUS from English to ASL (section 4) incorporates many key elements of the methodology of these prior translation projects.

2.3.1 Who translated the questionnaire, and how?

The methodology used in prior work for translating the survey instrument has varied. In some studies, a **team of experts** have performed the translation, sometimes as part of a multi-disciplinary team: For example, in the English-to-Turkish CSUQ project, a multidisciplinary research group of language professionals and reviewers were responsible for the translation. The group consisted of five native Turkish reviewers, two were bilingual reviewers, a native Turkish language expert and also usability expert who was responsible for coordinating the translation effort [12]. In the English-to-Persian SUS project [9], an ergonomics specialist conducted the initial translation. In [22], a **group of volunteers** suggested through online crowdsourcing how to translate the questions into German. An advantage of this approach is that it reduces the cost for researchers in conducting the translation, and it also provides a method by which questions could be translated into a wide variety of languages. However, the quality must be assessed, e.g., researchers in [22] had to edit the final version to improve the translation quality, subsequent to the crowdsourcing.

Section 4 will discuss how, in our project, we have asked a small **team of experts** to translate SUS into ASL.

2.3.2 How was the translation evaluated to determine if the original meaning was preserved?

There are several ways that the quality of the translation can be evaluated to determine whether it is fluent and whether it has preserved the meaning of the original English version.

In some projects, the team responsible for producing the translation conducts **multiple rounds of revision**, sometimes interleaved with rounds of user testing: For instance, in the English-to-Turkish CSUQ project [12], the quality of the translation was evaluated through a committee review process. After a draft of the translation to Turkish was developed from the original, the draft underwent a three-stage review process. This was aimed at modifying the Turkish version to preserve the meaning of the original English questions. Different participants were involved in each of the three stages. The nine members of the translation team independently evaluated the results after each round of testing and adjusted their translation to ensure that the meaning of the original questions were preserved. Each member of the research group reviewed the original and the revised versions of the Turkish SUS independently [12]. In the English-to-Persian SUS project [9], researchers conducted a study with 30 participants to identify linguistic problems in the translation; this pilot study led to the identification and integration of amendments into the final Persian SUS. In the English-to-Slovene SUS project [4], the translation process involved ten reviewers from the computer and natural sciences fields and three independent translators who were native Slovene speakers also fluent in English. The translation process was carried out in multiple stages, with members of the translation team considering comments from evaluators after each round. In the German-to-English SUS project [22], researchers edited the output of the crowdsourced translations written by volunteers online.

In some projects, a **back-translation procedure** is used to identify problems in the translated text: In this approach, after the team has produced their initial forward translation of a questionnaire from language "A" to language "B," then the researchers set up an evaluation study in which some new group of people translate each question back into the original language A. By comparing the original version of the survey (A) to this back-translated version (A-to-B-to-A), the researchers can determine if some meaning or concepts were lost during the original A-to-B translation phase. For example, in the English-to-Persian SUS project [9], two professional translators, who had lived in English-speaking countries, performed a back-translation after the original forward translation (English to Persian) had been carried out by the ergonomic specialist on the translation team. In the English-to-Slovene SUS project [4], the final round of revision included a back-translation procedure to look for any missing concepts.

Section 4 will discuss how, in our project, we have decided to use a translation process that incorporates both: **multiple rounds of revision** along with a **back-translation procedure** to identify any translated items that fail to preserve the meaning of the original.

2.3.3 How was the new version of the questionnaire evaluated to determine whether it preserved the useful psychometric properties of the original?

In addition to evaluating the text itself, researchers have evaluated whether key statistical characteristics of the questionnaire were preserved after the translation process. For example, in the English-to-Turkish CSUQ project [12], the researchers evaluated their translated version by using it in a usability study. Since CSUQ has a factor structure (i.e. it consists of several sub-scales, each based on a subset of questions), the researchers evaluated its **construct validity** by examining whether scores for clusters of items in the newly translated version were appropriately correlated.

In the English-to-Persian SUS project [9], a qualitative evaluation was also conducted involving a panel of ten experts to assess the **content validity** of the new instrument. Feedback from the expert group led to a revision of the translation of some items. Researchers also assessed their instrument through a study with 201 university students who evaluated a university food reservation system. The psychometric measures that were assessed based on these results included **construct validity** and **internal reliability** [9].

In the English-to-Slovene SUS project [4], a study involving 182 respondents (114 males, 86 females) was carried out using the new Slovene SUS to evaluate the usability of the Google Gmail website. The researchers evaluated their newly translated instrument's **internal reliability**, **criterion validity**, and **construct validity**.

Section 4 describes how we conducted a usability test with DHH participants responding to our new ASL version of SUS; we assessed **internal reliability** of our items using Cronbach's alpha [7]. We also asked participants to complete an additional usability questionnaire, which had previously been used in [2], to assess the **criterion validity** of our ASL translation of SUS.

2.4 Translating Surveys to ASL

Prior researchers have translated English surveys into ASL for health-related studies; although they had a focus on medical issues, these researchers' methodologies are also relevant to our work.

Health researchers have translated the Behavioral Risk Factor Surveillance System survey (a standardized survey conducted across the U.S.) into ASL using a **team of experts** consisting of bilingual individuals, with a mix of Deaf community members and health researchers [14]. The team strove for ASL with **meaning equivalence rather than word-for-word translation** and for ASL that would be understandable to a wide set of deaf individuals [14]. After the team produced an **ASL script** (a text file containing sequences of English words representing sequences of ASL signs, in appropriate ASL word order), a researcher not on the translation team **back-translated** the material to English. These were compared to the original English questions; in some cases, the translation team edited their ASL translations based on the results of this back-translation evaluation [14]. Finally, **native ASL signers** (people who grew up using ASL since early childhood) performed the ASL versions while being video recorded; an additional native ASL signer **"coach"** sat behind the camera, watching the script, to ensure that the performer followed the script that had been produced by the translation team.

Other researchers have translated the Multidimensional Health Locus of Control survey from English into ASL [24]. They convened a focus group of both bilingual members of the Deaf community who were **native ASL signers** and **ASL interpreters**. This first focus group produced an ASL translation for each English question item. Next, the researchers convened a second focus group (with similar membership composition) to produce an English **back-translation** for each ASL item; this second focus group also evaluated how close their English back-translation was to the original English version of the question items. In cases of divergence, the second focus group was asked to recommend a modification to the ASL version to better preserve the meaning of the original. During the back-translation evaluation, there were many cases (15 of 24) in which back-translation **lacked word-for-word equivalence** with the original English item, yet the focus group determined that the meaning was sufficiently preserved [24].

Section 4 will discuss how, in our project, the **team of experts** conducting the translation included **native ASL signers**. During translation, we created an **ASL script**, and during recording, a **coach** sat behind the camera while a **native signer** was recorded. During our evaluation, a group of **ASL interpreters** produced the **back-translations**, and when evaluating them, we considered **meaning equivalence rather than word-for-word translation**.

3. RESEARCH QUESTIONS

The goal of this study was to translate SUS, one of the most commonly used standardized questionnaires, from English into ASL to make it suitable for use with DHH users who prefer to communicate using ASL. Our challenge was to preserve the meaning of the questions and other desirable characteristics (section 2.2) so that results from our ASL instrument are comparable to results on the original English version. For convenience, we refer to our translated version as "ASL-SUS." In the remainder of this paper, we examine the following questions:

RQ1: Do the items in the ASL-SUS preserve the meaning of the original English SUS items, as measured through a back-translation study, in which ASL interpreters produce English back-translations and the meaning equivalence is compared to the original items?

RQ2: Deployed as the evaluation metric in a user study with DHH participants evaluating a university website, does ASL-SUS possess several key psychometric properties listed below?

a) **Internal Reliability**, as measured by Cronbach's alpha [7], indicating correlation among the set of ASL-SUS items

b) **Criterion Validity**, as measured by comparing the correlation between ASL-SUS and another measurement instrument collected concurrently from users: the adjective scale of [2]

4. RESEARCH METHODS

Our study included five phases: The first was the translation of SUS items into ASL. In phase two, a native ASL signer produced video recordings of the items, and the third phase consisted of back-translation and revision of the ASL videos. The fourth phase consisted of user studies with DHH users responding to the ASL-SUS items, to evaluate the quality of the translation and to assess the psychometric properties mentioned in RQ2 above. Finally, in phase five, we disseminate ASL-SUS to the research community. Details of each phase are discussed below.

4.1 Translation into ASL, Initial Revisions

During the first phase, a translation team was assembled, consisting of experts with skills in HCI and fluency in ASL and English:

- The first member was a doctoral student in computing, with several years of experience in the computing industry, who had completed two semester-long courses in HCI and was conducting research in HCI. This student was a fluent, native signer of ASL, born Deaf into a Deaf family of ASL-signers, and a graduate of a primary/middle/secondary school for the Deaf and from a university with instruction in ASL.
- The second member was a master's degree student in an HCI graduate program who had completed four semester-long courses in HCI, with experience conducting hundreds of hours of user testing, especially in-person studies with people who are DHH. This student was a fluent, native ASL signer who was born Deaf and used ASL since birth, and a graduate of a primary/middle/secondary school for the Deaf and from a university with instruction in ASL and English.
- The third member was a faculty member with a PhD in computing who publishes research in HCI and accessibility, and who had experience in corpus-based ASL computational linguistics research. This faculty member was a fluent signer who learned ASL as an adult, having completed 10 university courses in ASL, including 2 summer ASL immersion programs, and regularly uses ASL with DHH lab members.

The members of the team fell into specific roles: The faculty member suggested translation options and particular ASL linguistic structures (e.g. rhetorical questions, use of reference points in the signing space, etc.) to convey the original SUS question meaning. The doctoral student had final authority on the fluency of the ASL translations, given this student's "Deaf-of-Deaf" background and

native signing skills. The masters student identified terminology about technology likely understood by DHH participants in usability studies and when particular ASL options proposed by the doctoral student were so idiomatic as to be difficult to understand for a DHH participant who may possess non-native ASL fluency.

To begin the process, the translation team held several meetings to create draft concepts for how the SUS instructions[1] (presented at the beginning of the questionnaire) and the ten question items could be conveyed in ASL. During this time, the team found it useful to use a mobile phone video camera to record themselves producing alternatives and saving brief video clips representing alternatives. Multiple revisions were made for each SUS item, through this process, until a consensus set of ASL items were produced. The team avoided using an overly word-for-word transliteration of the original English items, to avoid producing non-fluent English-like signing videos. Instead, the team strove to preserve the overall meaning of each item, while using ASL structure. These casually-recorded videos were transcribed to produce an ASL script representing each item (see appendix A for more details).

4.2 Recording of High Quality ASL Videos

In phase two, high-definition videos were recorded of a native ASL signer performing each of the ASL-SUS items, in a video recording studio with professional-quality overhead lighting. The ASL signer wore solid-color black shirt (which contrasted with the signer's skin color, as shown in Figure 1), and a plain blue background (contrasting with both the skin and shirt color) was behind the signer. The master's degree student on the translation team acted as a coach, behind the camera, with a copy of the script, observing the ASL produced by the native signer being recorded to ensure that the performance matched the intended ASL translations.

4.3 Back-Translation, Final Revisions

In this third phase, we conducted a back-translation experiment to determine whether the translation preserved the meaning of the original. We recruited nine participants (5 female, 4 male) for this study, who were advanced students (3rd- or 4th-year) in a bachelor's degree program in ASL interpreting at Rochester Institute of Technology. Our rationale for recruiting interpreting students is that they regularly complete assignments in which they view and critique videos of ASL signing, and they have skills in discussing translation alternatives. The age of participants ranged from 20 to 26 (median 22). Three reported having deaf family members, including one participant who self-described as a Child of Deaf Adults (CODA). Aside from the CODA, who had used ASL since birth, the remaining participants had been using ASL for 3 to 8 years (mean 5.25). Participants were paid $40 for participating in this one-hour study, which was approved by the university IRB.

The participants were presented with each ASL-SUS video, and they were asked to write an English translation for each. Afterward, participants were shown the original English SUS, and they were asked to compare their translation to the original and to provide written feedback about any cases in which the meaning diverged. Participants were asked to offer suggestions about how the ASL video translation could be revised so that it would have been more successful at conveying the meaning of the original English item.

[1] In translating the instructions to ASL, additional explanatory content had to be added. For instance, the instructions had to explain that participants would need to watch *videos* of ASL and then circle items on paper, and the disagree/agree scale had to be explained in ASL (since the English questions would not appear on the paper). Further, the original SUS instructions mentioned the user's "reactions to the website today." Researchers generally change the word "website" to "device" or "software" as needed. To avoid producing multiple versions of our ASL videos, we needed the instructions to explain that the questions referred to the website, software, or device that the person had just used.

Figure 1. Example screenshots from videos of a native ASL signer performing some of the ASL-SUS items: (a) Excerpt from ASL-SUS for "I think that I would like to use this frequently," showing ASL sign FREQUENT, with cheeks-puffed-air-released-on-side facial expression to increase the degree of magnitude of the concept "frequently"; (b) Excerpt from ASL-SUS for "I found the system unnecessarily complex," showing ASL sign COMPLEX, with a negative-grimace facial expression to indicate bemused irritation.

The translation team (section 4.1) analyzed the back-translations and feedback comments from this study in order to identify any ASL videos for which the back-translations diverged from the meaning of the original English item or for which the participants had indicated suggestions of how the ASL video could have been improved. The translation team identified 2 items that needed to undergo major revisions based on the feedback, and 4 items that needed some small changes. Some examples of revisions made to our original ASL videos based on this feedback include:

- The initial translation (section 4.1) of SUS question 8 "I found the product very awkward to use" into ASL was as follows: I LOOK THIS$_{down}$, topic-eyebrow-raise{USE} AWKWARD. OVERWHELM (shrug). (Appendix A describes our notation used for ASL transcription.) Some participants indicated that the signer's performance of AWKWARD in the video was somewhat fast and difficult to perceive. Furthermore, several indicated that the use of OVERWHELM suggested that the fault lay with human user (who was not able to understand the technology), rather than the technology having some flaw. For instance, one back-translation was "This is over my head." Therefore, in the revised version of this item, the translation

team decided to omit the sign OVERWHELM, and the sign AWKWARD was slowed down (with a "disgust-tense" facial expression added) to improve intelligibility.

- Our original translation of SUS question 5 "I found the various functions in the product were well integrated" into ASL was as follows: I LOOK THIS$_{down}$, DO-DO$_{repetitive\text{-}circular\text{-}sweeping}$ cheek-puff{ MAINSTREAM-INTEGRATE } check-puff-side-release{ SMOOTH }. In the feedback comments, most participants indicated that they had not understood the signer's use of a repeated DO-DO sign performed in a circular sweeping arc motion to convey that the product does lots of things. For instance, one of the back-translations was "I can see all the pieces coming together." We decided to make use of the ASL sign ACTIVITY to more overtly refer to the many functions of the system[2]. The revised ASL version of this item was: I LOOK THIS$_{down}$, MANY ACTIVITY (signer sets up a list buoy[3], with rhythmic nodding when pointing to each item) I CAN USE TOGETHER FINE, WAVE-WOW.

We confirmed that our revisions had addressed the concerns raised in the back-translation study by evaluating the revised videos in a second back-translation study with 10 new participants (8 female, 2 male; ages 20 to 23; mean 9.4 years of ASL usage; all in the 3rd and 4th year in the ASL interpreting degree program).

4.4 Summative User Study Evaluation

To evaluate our revised ASL-SUS items, we conducted two studies in which DHH participants evaluate the usability of a university website. The only difference between the studies was the number of participants (30 in study #1, 10 in study #2) and whether they responded to the ASL-SUS (study #1) or the original English version (study #2), after they interacted with the website. Participants were paid $40 for participating in either one-hour study; both were approved by our university IRB.

4.4.1 Summative User Study #1 with ASL-SUS

In determining whether ASL signing participants could easily understand and respond to a SUS survey presented via ASL videos, we conducted a usability study with 30 participants (15 female, 15 male) from the Rochester Institute of Technology and surrounding community. Participant ages ranged from 19 to 27 (mean 22.5). Of the 30, 25 self-identified as Deaf, and 5, as Hard-of-Hearing. Twenty participants had used ASL since infancy, 6 since age 3, and the remainder by age 5. Eleven participants indicated that they had grown up in households with ASL signing parents.

Using a methodology similar to a prior study that had used SUS [6], we asked participants to access our university website and perform information-seeking tasks, without using a search engine:

- What events are happening on campus on March 9th, 2017?
- Find the textbook that you must buy for a given course.
- Find the phone number for the student employment office.
- Where is a particular building on the campus map?
- Find the latest issue of some university publication.
- Where in the library is "the Notebook" by Nicholas Sparks?
- What is on the menu today at the cafeteria?
- What are the opening hours for the academic support office?
- What's the phone number and office hours for the registrar?
- When does the freshman orientation event begin?

A fluent ASL signer conducted each experimental session in ASL, and the participant was handed a sheet of paper with the list of items above (written in English). Participants were asked to complete each task in sequence, and some tasks required up to five minutes. Most participants struggled with a few of the items on the list, with many commenting that they usually use the search engine to find information, rather than using the navigation bars on the website. Participants' computer screens were recorded during the session.

After completing the tasks, the participants were asked to view the ASL-SUS videos (consisting of an introductory video with instructions, followed by the ten Likert-type items). (Appendix A contains transcripts of the ASL-SUS items and a URL where the videos can be downloaded or viewed.) Participants indicated their answer choice for each Likert item on a piece of paper that contained a list of numbers from 1 to 10 arranged vertically, with a set of five checkboxes next to each. The leftmost checkbox was labeled "strongly disagree," and the rightmost, "strongly agree." No other English questions or instructions appeared on the paper.

In order to evaluate the criterion validity of ASL-SUS, it is useful to collect responses from participants on a second pre-existing usability instrument. In [2], researchers wanted to identify English adjectives that corresponded to various SUS scores [2]; so, they asked 964 English-speaking participants (mean age 40.4; 474 female, 490 male) to evaluate a variety of websites and software, across several studies. The participants responded to the original English version of the SUS questionnaire, and in addition, they responded to a novel seven-point "adjective scale" designed by the researchers. Participants were asked to select a word that indicated how user-friendly they believed the website or software to be. The scale consisted of adjectives, scored as ordinal values: Worst Imaginable (1), Awful (2), Poor (3), OK (4), Good (5), Excellent (6), Best Imaginable (7). We decided to ask our participants to respond to this adjective scale in our study: since this instrument had previously been compared to the original English SUS, we could perform a similar analysis of our new ASL-SUS. Since our participants were fluent ASL signers, we provided an ASL video translation of the Adjective Scale user-friendliness question and the answer choice levels (see Appendix A), but the paper answer-sheet presented the seven options using the original English adjectives.

4.4.2 Comparison Study #2 with English SUS

While there are many published results and norms available for English-speaking users responding to original English SUS items, we wanted to gather a small set of response data from additional DHH participants evaluating the university website. Section 5 discusses how this comparative data is used to evaluate the ASL-SUS. In this study #2, 10 DHH participants (4 female, 6 male) were recruited from the university campus and surrounding community. Participant ages ranged from 20 to 27 (mean 22.4). Of the 10, 5 self-identified as Deaf, and 5, as Hard-of-Hearing. Three had used ASL since infancy, 2 since age 4, 2 since age 7, and the remainder by age 16. Five participants indicated that they had grown up in households with ASL signing parents. The procedure and tasks in this study were identical to those in study #1, with one difference: Instead of responding to the ASL-SUS, these participants

[2] While some ASL signers use a specific sign for "function," the native ASL signers on the translation team felt that this was an English-influenced sign, using an "F" handshape initialization, and it may not be widely understood by signers less familiar with computer jargon.

[3] In ASL, a "list buoy" refers to a linguistic construction in which the signer raises the non-dominant hand into the signing space with some number of fingers extended, to represent a list of items that is being communicated. The signer generally points to each finger when referring to or introducing each item in the list. Here, the list buoy is used in a non-specific sense, to indicate that there are a multitude of functions.

responded to the original English SUS items on a paper questionnaire.

4.5 Dissemination of ASL-SUS

In the final phase of our study, we have disseminated our new ASL-SUS, to share this resource with the research community. Appendix A provides a transcript of each item, and the final version of the ASL videos and PDFs of the paper answer sheets are available at http://latlab.ist.rit.edu/assets2017sus

To provide the research community with additional information about participant responses on these new instruments, we share the raw numerical response data from study #1 and #2 (on ASL-SUS, English SUS, and Adjective Scale) in Appendix B.

5. Results

This section analyzes the results from our evaluation of ASL-SUS, with a focus on addressing the research questions in section 3.

Our first research question (**RQ1**) was whether we had preserved the meaning of the original English SUS during the translation process used to produce ASL-SUS. This issue was primarily examined in a formative manner, during the translation and back-translation evaluation process. To ensure that meaning was preserved during translation, we used several commonly accepted methodologies: a team of experts produced the translation, multiple rounds of revision occurred, a back-translation evaluation was conducted prior to the final revision, and native ASL signers participated in the translation process. Informally, we note that the mean scores for ASL-SUS in study #1 (52.25) and the original English SUS in study #2 (50.5) were quite similar; this is a reasonable result since DHH participants evaluated an identical website and performed identical search tasks in both studies.

Although not a formal research question identified in this study, we were also interested in the overall usability of the ASL-SUS, i.e., whether DHH participants found any of the videos confusing or difficult to understand. During summative study #1 (section 4.4.1), we invited participants to indicate if they were unsure of the meaning of any ASL-SUS videos or had difficulty in responding to any questions. At the end of the study, participants wrote feedback comments about their experience participating in the study. None of the 30 participants indicated that they experienced difficulty in understanding or responding to the ASL-SUS questions.

In regard to **RQ2**, sections 5.1 and 5.2 analyze the results from studies #1 and #2 to examine two key psychometric properties: **internal reliability** and **criterion validity**.

5.1 Internal Reliability

To evaluate the internal reliability of ASL-SUS, i.e. whether all items contribute to the overall score, we calculated Cronbach's alpha [7]. Since even-numbered items on SUS are reverse-scored (they have negative polarity), we inverted those items. For responses from the study #1 participants, the alpha for ASL-SUS was 0.69. Generally, alpha scores of 0.7 and above are considered acceptable [18]; so, our ASL-SUS is on the borderline.

To assess the internal reliability of our new ASL-SUS, we compare this alpha value above with two other alpha values:

- In a **prior study** examining the internal reliability of the original English SUS, researchers reported alpha scores of 0.91, based on responses from 2,324 hearing participants [3].
- In our **study #2**, 10 DHH participants responded to the original English SUS, while evaluating an identical university website as in study #1; we calculated an alpha value of 0.79.

Given the alpha value of 0.69 for ASL-SUS in study #1 (and the relatively similar alpha values for English SUS in study #2, also with DHH participants), we conclude there is evidence for the internal reliability of the ASL-SUS. However, the alpha for ASL-SUS was lower than published values for the original English SUS on response data from a large number of hearing participants.

5.2 Criterion Validity

One method of evaluating criterion validity is to determine how convergent an instrument is with other instruments that measure the same abstract concept. Section 4.4 described how our DHH participants, in addition to responding to ASL-SUS, also responded to the Adjective Scale from [2]. To assess whether the scores from each participant on the ASL-SUS were correlated to the scores from that participant on the Adjective Scale, we calculated the Pearson's coefficient $r=0.684$, $p<0.001$. This result indicated a significant correlation between participants' responses to these instruments.

To assess the criterion validity of ASL-SUS, we compare this correlation above with two other correlations:

- **Previously Published Correlation between English SUS and Adjective Scale:** In [2], the authors calculated the correlation between their participants' English SUS scores and their Adjective Scale scores: Pearson's $r=0.822$, $p<0.01$. We performed a Fisher's *r*-to-*z* transformation to determine whether this ASL-SUS-to-Adjective-Scale correlation from study #1 was significantly different than the English-SUS-to-Adjective-Scale correlation reported in [2]; no significant difference was observed ($p=0.09$, $z=1.67$).
- **Study #2 Correlation between English SUS and Adjective Scale:** The participants who provided the response data in [2] consisted almost entirely of hearing participants. To compare our study #1 correlation above with data from a more similar group of users, we can examine the data from study #2, in which 10 DHH participants responded to the original English SUS and to the Adjective Scale. We calculated the correlation between SUS and Adjective Scale for these users: Pearson's $r=0.633$, $p<0.5$. To compare our study #1 results and study #2 results, a Fisher's *r*-to-*z* transformation was performed to determine whether our study #1 ASL-SUS-to-Adjective-Scale correlation was significantly different than the study #2 English-SUS-to-Adjective-Scale correlation; no significant difference was observed ($p=0.42$, $z=0.21$).

Thus, we conclude that there is strong evidence that the ASL-SUS has comparable criterion validity to the original English SUS, using the adjective scale of [2] as our criterion.

6. CONCLUSIONS AND FUTURE WORK

The specific goal of this project was to create and evaluate an ASL translation of the popular SUS questionnaire, commonly used by HCI researchers and practitioners to quickly evaluate the usability of websites or software. We investigated (sections 2.3.1 and 2.4) and utilized (sections 4.1, 4.2, and 4.3) accepted methodologies for formally translating a pre-existing English instrument into ASL.

We also investigated (sections 2.3.2 and 2.3.3) and utilized (sections 4.4 and 5) accepted methodologies for evaluating whether a translated instrument has preserved the meaning of the original version and whether it possesses various psychometric properties. Our results provide strong evidence of the criterion validity of ASL-SUS (as related to the adjective scale of [2]) and moderate evidence of its internal reliability (based on a Cronbach's alpha calculation). In summary, the results of our summative evaluation

of ASL-SUS indicate that this instrument is suitable for use by researchers who wish to evaluate software or websites through studies with DHH participants.

The broader goal of this study has been to facilitate the participation of people who are DHH in HCI research studies or in commercial usability testing by HCI practitioners, including during studies that primarily include hearing participants. By making it easier for HCI researchers to provide DHH users with an understandable version of this common usability questionnaire (SUS), we hope to reduce a barrier for including DHH participants.

6.1 Limitations

While providing ASL versions of standard surveys may address the issue of lower English literacy rates among many deaf individuals that would make English instruments difficult to use, this study has not addressed other potential barriers, including: providing access to low-literacy DHH individuals with weaker ASL skills, helping researchers find and recruit DHH individuals for studies, or supporting communication of task instructions or other information for DHH users that is necessary for participation in the study.

While we evaluated several psychometric properties, there are additional properties that should be evaluated in subsequent studies, e.g. external reliability (i.e. whether someone re-taking ASL-SUS on another occasion obtains a similar score). Furthermore, our analysis was based on data collected from DHH participants who evaluated a single website; it would be valuable to obtain ASL-SUS scores for a wider variety of products to serve as a broader basis for subsequent psychometric evaluations – to determine if ASL-SUS behaves similarly to original SUS in a variety of evaluation circumstances. Additionally, the DHH participants in our study were mostly university students between ages 20 and 27; for future psychometric evaluation, we should record responses from a wider age range of DHH participants.

6.2 Future Work

While an in-person evaluation study with 30 DHH users is reasonable in size, given comparable prior published accessibility studies with DHH users [15, 16], some recommend [12] recruiting a total number of participants equal to five times the number of items on an instrument that is being psychometrically validated, i.e. 50 in the case of SUS. Our lab regularly conducts usability studies with ASL users, and we intend to continue to make use of ASL-SUS, to enlarge our collection of response data – and to replicate our psychometric evaluation of ASL-SUS. As we collect this additional data, we will address several of the limitations outlined above (e.g. evaluating different technology artifacts, and including a wider age range of participants), and we intend to evaluate additional psychometric properties of ASL-SUS in future work. For instance, a prior study [2] concluded that SUS had a single factor structure (based on an examination of the results of a single-factor analysis), yet researchers in [21] discuss how specifically investigating multi-factor solutions led them to conclude that SUS actually consists of two factors: Learnability (based on the responses to items 4 and 10) and Usability (based on the responses to the remaining items). We would like to evaluate the construct validity of ASL-SUS, by determining whether this factor structure had been preserved during the process of translating the items to ASL. However, our study #1 only included 30 participants. In future work, we plan on collecting additional response data and then conducting a common factor analysis with varimax rotation, as in [21], to compare the factor structure of ASL-SUS to the factor structure of the original SUS.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under award number 1400802 and 1462280, and it has been supported by a Google Faculty Research Award and by the National Technical Institute for the Deaf (NTID).

8. REFERENCES

[1] American Educational Research Association, Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing.* AERA, Washington, DC.

[2] Bangor, A., Miller., J., Kortum, P. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies* 4(3), 114-123.

[3] Bangor, A., Kortum, P., Miller, J.T. 2008. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction.* 24, 574-594

[4] Blažica, B., Lewis, J.R. 2015. A Slovene translation of the system usability scale: the sus-si. *Int J of Hum-Comput Interact* 31(2), 112-117. DOI: http://dx.doi.org/10.1080/10447318.2014.986634

[5] Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

[6] Chaparro, B.S. 2008. Usability evaluation of a university portal website. *Usability News*. (October 2008) Retrieved May 1, 2017 from http://usabilitynews.org/usability-evaluation-of-a-university-portal-website/

[7] Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 297–334. DOI: http://dx.doi.org/1010.1007/BF02310555

[8] Da Silva Alves, A., Ferreira, S. B., De Oliveira, V.S., Da Silva, D.S. 2012. Evaluation of potential communication breakdowns in the interaction of the deaf in corporate information systems on the web. *Procedia Comput Sci* 14, 234-244. DOI: http://dx.doi.org/10.1016/j.procs.2012.10.027

[9] Dianat, I., Ghanbari, Z., Asghari Jafarabadi, M. 2014. Psychometric properties of the Persian language version of the system usability scale. *Health Promotion Perspectives* 4(1), 82-89. DOI: http://dx.doi.org/10.5681/hpp.2014.011

[10] Dilsizian, M., Tang, Z., Metaxas, D., Huenerfauth, M., Neidle, C. 2016. The importance of 3D motion trajectories for computer-based sign recognition. In *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, The 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia. ELRA, Paris.

[11] Elliot, L., Stinson, M., Mallory, J., Easton, D., Huenerfauth M. 2016. Deaf and hard of hearing individuals' perceptions of communication with hearing colleagues in small groups. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. ACM, Reno, NV, USA, 271-272. DOI: http://dx.doi.org/10.1145/2982142.2982198

[12] Erdinç, O., Lewis, J.R. 2013. Psychometric evaluation of the T-CSUQ: the Turkish version of the computer system usability questionnaire. *Int J of Hum-Comput Interact* 29(5), 319-326. http://dx.doi.org/10.1080/10447318.2012.711702

[13] Geisinger, K.F. 1994. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment* 6(4), 304–312.

[14] Graybill, P., Aggas, J., Dean, R. K., Demers, S., Finigan, E., Pollard, R.Q. 2010. A community-participatory approach to adapting survey items for deaf individuals and American Sign Language. *Field Methods* 22(4), 1-20.

[15] Huenerfauth, M., Kacorri, H. 2014. Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, The 9th International Conference on Language Resources and Evaluation (LREC 2014),* Reykjavik, Iceland. ELRA, Paris.

[16] Kacorri, H., Huenerfauth, M., Ebling, S., Patel, K., Menzies, K., Willard, M. 2017. Regression analysis of demographic and technology-experience factors influencing acceptance of sign language animation. *ACM Trans. Access. Comput.* 10, 1, Article 3 (April 2017), 33 pages. DOI: http://dx.doi.org/10.1145/3046787

[17] Kawas, S., Karalis, G., Wen, T., Ladner, R.E. 2016. Improving real-time captioning experiences for deaf and hard of hearing students. In *Proc. 18th International ACM SIGACCESS Conf. on Computers and Accessibility.* ACM, New York, NY, USA, 15-23. DOI: http://dx.doi.org/10.1145/2982142.2982164

[18] Landauer, T.K. 1997. Behavioral research methods in human-computer interaction. In M. Helander, T. K. Landauer, and P. Prabhu (eds.) *Handbook of human–computer interaction, 2nd ed.*, 203–227. Elsevier, Amsterdam

[19] Lazar, J., Feng, J.H., and Hochheiser, H. 2017. *Research Methods in Human-Computer Interaction, Second Edition.* Morgan Kaufmann, San Francisco, CA, USA.

[20] Lewis, J.R. 1995. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Int'l Journal of Human–Computer Interaction,* 7, 57–78.

[21] Lewis, J.R., Sauro, J. 2009. The factor structure of the system usability scale. In *Proc. of Human Computer Interaction International conference (HCII 2009)*, San Diego CA.

[22] Lohmann, K., Schäffer, J. 2013. System usability scale (SUS)-An improved German translation of the questionnaire. (September 2013) Retrieved May 1, 2017 from http://minds.coremedia.com/2013/09/18/sus-scale-an-improved-german-translation-questionnaire/

[23] Mitchell, R., Young, T., Bachleda, B., and Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies* 6(3):306-335.

[24] Samady, W., Sadler, G.R., Nakaji, M., Malcarne, V.L. 2008. Translation of the multidimensional health locus of control scales for users of American Sign Language. *Public Health Nursing* 25(5), 480-489. DOI: http://dx.doi.org/10.1111/j.1525-1446.2008.00732.x

[25] Sauro, J., Lewis, J.R. 2012. *Quantifying the user experience: Practical statistics for user research.* Elsevier, Amsterdam.

[26] Schnepp, J., Shiver, B. 2011. Improving deaf accessibility in remote usability testing. In *Proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '11)*. ACM, New York, NY, USA, 255-256. DOI: http://dx.doi.org/10.1145/2049536.2049594

[27] Sears, A., Hanson, V. 2011. Representing users in accessibility research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11).* ACM, New York, NY, USA, 2235-2238. DOI: http://dx.doi.org/10.1145/1978942.1979268

[28] Tran, J.J., Johnson, T.W., Kim, J., Rodriguez, R., Yin, S., Riskin, E.A., Ladner, R.E., Wobbrock, J.O. 2010. A web-based user survey for evaluating power saving strategies for deaf users of mobileASL. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility.* ACM, New York, NY, USA, 115-122. DOI: http://dx.doi.org/10.1145/1878803.1878825

[29] Traxler, C. 2000. The Stanford achievement test, 9th ed: national norming and performance standards for deaf & hard-of-hearing students. *J Deaf Stud & Deaf Educ* 5(4): 337-348.

9. APPENDIX A: ASL TRANSLATIONS

Videos of the ASL version of the SUS question items appear online at this URL: http://latlab.ist.rit.edu/assets2017sus

A gloss transcription of each ASL video, using English words to informally represent ASL signs, appears below for the reader's convenience. We include below some notes on notational conventions used in this appendix and in section 4.3: There are a wide variety of notational conventions used by ASL linguists and interpreters to informally represent ASL signing in a written form; for this paper, we have used the following conventions:

- Capital letters will be used to represent individual signs. When multiple English words are being used to represent a single ASL sign, then hyphens will be used, e.g., DON'T-MIND is a single sign in ASL (i.e. right index finger sliding off the nose and then pointing forward away from the body).
- Lowercase letters and braces (i.e. "{ }") indicate non-manual aspects of the performance and the span of words during which they occur, e.g. disgust-face-without-head-shake{ DIFFERENT} indicates that the signer performs the sign DIFFERENT while simultaneously producing a negative disgust facial expression (without any side-to-side head shaking, which often co-occurs with this facial expression).
- Additional gestural aspects of the performance that are difficult to convey using the above notations are simply indicated with lowercase words in parentheses, e.g., (shrug).

9.1 ASL Translation of SUS

Prior to the 10 question items, the SUS customarily begins with brief instructions. The ASL translation below differs in content from the English text since the participant must be instructed to watch videos in the ASL version. In addition, the signer explains the Likert-style response scale for the question items, with options: strongly disagree, disagree, neutral, agree, and strongly agree.

Introduction: *English:* "Instructions: For each of the following statements, mark one box that best describes your reactions to the website today." *ASL Transcript:* HELLO, rhetorical{ THIS QUESTION WHAT}, SEE THIS PAPER, HAVE SENTENCE. I WILL SIGN EACH SENTENCE. rhetorical{ YOU DO-DO } WATCH POINT SENTENCE, WATCH, WONDER YOU AGREE OR DISAGREE, wh-question{ WHICH }. yes-no-question{ REMEMBER BEFORE }, YOU PLAY $\text{THIS}_{\text{down}}$ THING, VARIOUS, MAYBE PROGRAM, MAYBE WEBSITE, MAYBE ELECTRIC, DIFFERENT. PLAYING $\text{THIS}_{\text{down}}$ TEST. NOW, YOU DREAM-BUBBLE-IMAGINE QUESTION ASL. rhetorical{ AGREE OR DISAGREE, HOW ANSWER}. SEE POINT BOX 1 2 3 4 5, (nodding). rhetorical{ 1 WHAT } STRONGLY DISAGREE. rhetorical{ 2 WHAT } OKAY DISAGREE. rhetorical{ 3 WHAT } NEUTRAL. rhetorical{ 4 WHAT } OKAY AGREE. rhetorical{ 5 WHAT } STRONGLY AGREE. YOU OBSERVE, PICK ONE, CIRCLE ANSWER.

Note: The ASL signer in the video introduction above pointed to a location in the lower region of the signing space to represent the concept of the product, website, or software that the participant had just used. In the transcriptions below, whenever the signer refers to the "product," he points to this region in the signing space again. This avoids the redundancy of the signer repeating the phrase "software, website, or product" for every question item below. This pointing sign is glossed using $THIS_{down}$ in the transcriptions below.

Question 1: *English:* "I think that I would like to use this product frequently." *ASL Transcript:* I LOOK $THIS_{down}$, head-nod-with-mouth-morpheme-cheek-puff{ DON'T-MIND USE OFTEN }

Question 2: *English:* "I found the product unnecessarily complex." *ASL Transcript:* I LOOK $THIS_{down}$, negative-grimace{ WAVE-WOW COMPLEX }

Question 3: *English:* "I thought the product was easy to use." ASL Transcript: I LOOK $THIS_{down}$, head-nod-with-mouth-morpheme-cheek-puff{ EASY USE UNDERSTAND }

Question 4: *English:* "I think that I would need the support of a technical person to be able to use this product. " ASL Transcript: I DREAM-BUBBLE-IMAGINE $THIS_{down}$ USE, I NEED TECHNOLOGY $HELP_{me}$ (emphatic-nod)

Question 5: *English:* "I found the various functions in the product were well integrated." *ASL Transcript:* I LOOK $THIS_{down}$, MANY ACTIVITY (list buoy using the left hand with four fingers extended to indicate list items, using right index finger to point to each item one-by-one, with rhythmic nodding when pointing to each item) I CAN USE TOGETHER FINE, WAVE-WOW.

Question 6: *English:* "I thought there was too much inconsistency in this product." *ASL Transcript:* I LOOK $THIS_{down}$, MANY ACTIVITY (list buoy using the left hand with four fingers extended to indicate list items, using right index finger to point to each item one-by-one, with rhythmic nodding when pointing to each item) disgust-face-without-headshake{ DIFFERENT } disgust-face-with-negative-headshake{ NOT SAME }

Question 7: *English:* "I imagine that most people would learn to use this product very quickly." *ASL Transcript:* I DREAM-BUBBLE-IMAGINE LOOK $THIS_{down}$, LEARN USE FAST (shrug)

Question 8: *English:* "I found the product very awkward to use." *ASL Transcript:* I LOOK $THIS_{down}$, I USE disgust-face-without-head-shake-with-tense-grimace{ AWKWARD }

Question 9: *English:* "I felt very confident using the product." *ASL Transcript:* I LOOK $THIS_{down}$, I USE happy-positive-face{ CONFIDENT }

Question 10: *English:* "I needed to learn a lot of things before I could get going with this product." *ASL Transcript:* I LOOK $THIS_{down}$, WAVE-WOW I MUST LEARN MANY MANY, I FINALLY USE

9.2 ASL Translation of Adjective Scale

The adjective scale was designed in [2]. The translation team decided that the terms "user-friendliness" and "product" needed some expansion in ASL to convey the meaning, e.g. to suggest that the person might have seen a website, a computer program, etc.

English: "Overall, I would rate the user-friendliness of this product as: worst imaginable, awful, poor, OK, good, excellent, best imaginable." *ASL Transcript:* yes-no-question{ REMEMBER BEFORE YOU PLAY THIS THING } VARIOUS, MAYBE PROGRAM, MAYBE WEBSITE, MAYBE ELECTRIC. PLAY, TEST, (nodding). NOW, I CURIOUS, WHAT YOUR OPINION $THIS_{down}$. rhetorical-question{ HOW EXPLAIN THIS } yes-no-question{ SEE BOX (points to first checkbox on paper sheet) } YOU OBSERVE, I EXPLAIN. FIRST, RIGHT SIDE BOX, NO GOOD disgust-face{ WORST } NEXT, disgust-face{ AWFUL } NEXT, disgust-face{ POOR } NEXT, grimace{ OK } NEXT, shrug{ GOOD } NEXT, positive-face{ EXCELLENT } NEXT, positive-face{ CHAMP-BEST } NOW, OBSERVE $THIS_{down}$, YOU PICK ONE, CIRCLE.

10. APPENDIX B: USER STUDY DATA

This appendix contains the raw data collected in study #1 and study #2, described in section 4.4. The Tables 1 and 2 contain the following columns: "Partic. ID #" (a code number used to refer to each participant), "Q1" to "Q10" (the raw scores from participant responses to individual Likert-type items, with 1 indicating Strongly Disagree and 5 indicating Strongly Agree), the ASL-SUS score based on these ten responses, and the "Adj. Scale" indicating the numerical values corresponding to the adjective choice of the participant on this item (scoring details appear in section 4.4.1). Because half of the individual items on SUS have negative polarity, overall scores are calculated according to this formula, from [5]:

$$\mathbf{SUS} = 2.5 \cdot ((\mathbf{Q1}\text{-}1) + (5\text{-}\mathbf{Q2}) + (\mathbf{Q3}\text{-}1) + (5\text{-}\mathbf{Q4}) + (\mathbf{Q5}\text{-}1) + (5\text{-}\mathbf{Q6}) + (\mathbf{Q7}\text{-}1) + (5\text{-}\mathbf{Q8}) + (\mathbf{Q9}\text{-}1) + (5\text{-}\mathbf{Q10}))$$

Table 1: Raw data from user study #1 with ASL-SUS.

Partic. ID #	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	ASL-SUS	Adj. Scale
1	3	2	3	2	5	4	4	1	4	2	70	5
2	3	1	5	4	3	4	5	2	5	3	67.5	6
3	4	4	2	3	4	4	2	4	2	4	37.5	4
4	3	2	3	1	3	3	5	4	3	3	60	5
5	3	4	4	5	4	5	5	3	4	4	47.5	5
6	3	4	2	4	4	4	4	2	3	2	50	5
7	5	4	4	4	4	4	3	3	4	4	52.5	5
8	3	2	3	4	4	2	3	2	5	5	57.5	5
9	4	3	2	3	4	2	3	3	4	3	57.5	5
10	3	2	4	1	3	5	4	2	4	3	62.5	5
11	5	2	4	2	4	3	2	2	4	3	67.5	6
12	3	2	3	4	4	4	4	3	3	5	47.5	7
13	2	4	2	3	4	4	2	3	2	4	35	4
14	1	4	2	2	2	4	1	2	2	4	30	1
15	2	4	2	3	3	4	2	4	1	5	25	2
16	4	2	4	2	2	4	4	2	4	4	60	5
17	3	4	2	3	3	4	2	3	4	3	42.5	5
18	4	3	4	3	2	4	4	3	3	3	52.5	5
19	5	3	4	4	5	4	3	3	5	4	60	5
20	1	4	4	5	5	5	5	5	5	5	40	6
21	1	3	3	3	3	3	2	5	3	3	37.5	4
22	2	4	3	4	2	2	1	3	3	5	32.5	4
23	3	2	4	3	5	5	4	4	4	4	55	5
24	4	3	2	4	4	4	3	5	3	5	37.5	4
25	5	3	3	5	4	2	2	1	3	3	57.5	5
26	4	2	3	2	4	1	5	1	4	4	75	6
27	4	4	3	4	4	3	4	4	3	4	47.5	5
28	5	2	4	2	4	1	3	2	4	2	77.5	6
29	3	2	4	2	4	2	3	3	3	2	65	5
30	5	4	5	4	5	5	5	4	5	4	60	6

Table 2: Raw data from user study #2 with English SUS.

Partic. ID #	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	English SUS	Adj. Scale
31	4	5	5	1	5	2	3	2	4	1	75	6
32	3	3	4	2	5	3	4	2	4	2	70	5
33	4	3	4	4	3	4	3	3	3	3	50	4
34	2	1	3	1	4	1	5	3	4	5	67.5	4
35	2	3	3	5	4	3	4	5	2	4	37.5	4
36	4	3	4	2	4	3	3	3	4	2	65	6
37	2	5	3	2	3	4	1	5	4	1	40	4
38	1	5	4	1	1	5	3	4	4	4	35	3
39	3	4	2	5	3	3	2	5	3	3	32.5	5
40	2	4	2	4	3	4	3	3	3	5	32.5	4