Regression Analysis of Demographic and Technology-Experience Factors Influencing Acceptance of Sign Language Animation

HERNISA KACORRI, Carnegie Mellon University
MATT HUENERFAUTH, Rochester Institute of Technology
SARAH EBLING, University of Zurich
KASMIRA PATEL, KELLIE MENZIES, and MACKENZIE WILLARD,
Rochester Institute of Technology

Software for automating the creation of linguistically accurate and natural-looking animations of American Sign Language (ASL) could increase information accessibility for many people who are deaf. As compared to recording and updating videos of human ASL signers, technology for automatically producing animation from an easy-to-update script would make maintaining ASL content on websites more efficient. Most sign language animation researchers evaluate their systems by collecting subjective judgments and comprehension-question responses from deaf participants. Through a survey ($N = 62$) and multiple-regression analysis, we identified relationships between (a) demographic and technology-experience characteristics of participants and (b) the subjective and objective scores collected from them during the evaluation of sign language animation systems. These relationships were experimentally verified in a subsequent user study with 57 participants, which demonstrated that specific subpopulations have higher comprehension or subjective scores when viewing sign language animations in an evaluation study. This finding indicates that researchers should collect and report a set of specific characteristics about participants in any publications describing evaluation studies of their technology, a practice that is not yet currently standard among researchers working in this field. In addition to investigating this relationship between participant characteristics and study results, we have also released our survey questions in ASL and English that can be used to measure these participant characteristics, to encourage reporting of such data in future studies. Such reporting would enable researchers in the field to better interpret and compare results between studies with different participant pools.

CCS Concepts: • **Human-centered computing → User studies; Empirical studies in accessibility**

Additional Key Words and Phrases: Accessibility technology for people who are deaf, American Sign Language, animation, user study

ACM Reference Format:
Hernisa Kacorri, Matt Huenerfauth, Sarah Ebling, Kasmira Patel, Kellie Menzies, and Mackenzie Willard. 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

This material is based upon work supported by the National Science Foundation under awards No. 1065009 and No. 1506786, and by the Swiss Science Foundation, under award NO. 155263.

Authors' addresses: H. Kacorri, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA; email: hkacorri@cmu.edu; M. Huenerfauth, K. Patel, K. Menzies, and M. Willard, Rochester Institute of Technology, 152 Lomb Memorial Drive, Rochester, NY 14623, USA; emails: {matt.huenerfauth, kxp3157, ksm2428, mtw2935}@rit.edu; S. Ebling, University of Zurich, Plattenstrasse 47, 8032 Zurich, Switzerland; email: ebling@cl.uzh.ch.

DOI: http://dx.doi.org/10.1145/3046787

1. INTRODUCTION

With the increasing importance of the Internet for commerce, communication, education, and social networking, gaining access to online media and websites has become essential for full participation in modern society. The vast majority of this information online is in the form of written language text, which is not fully accessible to many groups of users. For instance, many people prefer to receive information content in the form of sign language, especially individuals who identify as Deaf.[1] In the U.S., over 500,000 people use American Sign Language (ASL) as a primary language [Mitchell et al. 2006].

Beyond this language preference, there are also trends in written language literacy that are important to consider: Due to reduced exposure to language during childhood or other educational circumstances, many people who are deaf and hard-of-hearing have lower levels of written language literacy. For example, in the U.S., standardized testing of high school graduates (secondary school, age 18+) has found that the median literacy rate of deaf high school graduates is at the 4th-grade level [Traxler 2000]. (Students in the 4th grade in the U.S. are typically age 10.) There are significant linguistic differences between English and ASL; therefore, it is possible to be fluent in one language but not the other. Thus, if websites or online media sources were able to provide information in the form of ASL, then this content would be more accessible to users with lower English literacy.

Many companies, organizations, and governments provide information content on their websites in multiple languages, in order to reach a diverse and global audience; so, it is reasonable to wonder why there are few websites that provide significant content in the form of sign language. One consideration is that there is no writing system for ASL that is in common use among the Deaf community; so, ASL content cannot be provided online in a text-based form. While it is seemingly simple to video-record someone performing ASL and post it on a website, the difficulty arises when attempting to efficiently maintain such content. When the information must be updated, the organization must rerecord the message and post the new video online. Beyond this maintenance issue, video-based solutions do not enable just-in-time generation of website content from a user query.

For this reason, several international research groups (e.g., Hayward et al. [2010], Jennings et al. [2010], Kennaway et al. [2007], Kipp et al. [2011], and Verlinden et al. [2001]) have investigated software to automatically synthesize accurate animations of a virtual human performing sign language, with the input to this software being an easy-to-update script of the message. (The script could be authored by someone knowledgeable of sign language—or possibly produced through some automated process.) A major technical challenge for these researchers is how to automatically select the details of the animations so that they are linguistically accurate, easily understandable, and acceptable to users. Researchers typically evaluate their software by automatically generating some animations using their software, conducting an experiment where deaf participants view and evaluate the animations, and comparing the scores of animations produced using the software (to some baselines or to animations produced by prior versions of the software).

Unfortunately, there is limited consensus about the set of demographic data that should be reported about the participants in these studies. Thus, it is difficult to compare the results across different studies because some of the variation in comprehension or subjective evaluation scores that is reported may be explained by the demographic characteristics of the particular participants in that study, rather than by true differences in the quality of the animations being evaluated. The goal of our research is to

[1] We follow the widely held convention of using the capitalized term "Deaf" to refer to people who identify as members of the Deaf Community or Deaf Culture, and we use "deaf" as a more general term.

examine whether demographic and technology-experience variables are predictors of participants' responses to (a) subjective measures of animation quality and (b) objective measures of comprehension of the content.

This article is an extended version of a paper originally presented at the 2015 ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'15) [Kacorri et al. 2015], which presented a study in which ASL signers were shown ASL animations (using a variety of avatars) and were asked questions of type (a) and (b). In addition, participants were asked questions about (i) demographic characteristics and (ii) their technology experience/attitudes. Multiple-regression analysis was used to determine whether variables (i) and (ii) relate to participants' responses (a) and (b).

This article contains additional analysis about the demographic and technology-experience characteristics of the participants in the original study (Sections 5.1 and 5.2), a comparison of how much variance in the responses of participants in the study were due to the differences in the animation stimuli themselves (Section 6.2), and more information about the unstructured text responses of participants in the study (Section 8). Most substantially, this article presents a follow-up study with 57 participants (Section 7) in which trends identified in the original study were experimentally evaluated, to enable us to evaluate the statistical significance of the relationships between user characteristics and their comprehension and subjective scores during evaluations of these animations. Finally, this article includes the English text and ASL video versions of the questions from our demographic and technology-experience survey that we recommend for use by future sign language animation researchers; these materials are available in the online appendix to this article available in the ACM Digital Library.

The remainder of this article is organized as follows: Section 2 presents related work on demographics and evaluation studies. Section 3 describes how demographic and technology-experience information was gathered from study participants, and Section 4 describes how comprehension scores and subjective scores were collected. Sections 5 and 6 describe the methodology and results of our initial study, and Section 7 presents a follow-up experimental study used to evaluate the findings from the first study. Section 8 describes feedback comments from participants, and finally, Section 9 summarizes our conclusions and future directions.

2. RELATED WORK

In prior research studies that have evaluated the quality of sign language animation technology, researchers have often asked human participants to respond to subjective questions about the animation output and, less frequently, to answer comprehension questions about the information content of those animations. The degree to which these researchers have considered or reported on the demographic or attitudinal characteristics of their participants has varied widely. This section will survey prior literature in this area, with a focus on studies that have been conducted with deaf participants in the context of evaluating sign language animations (Section 2.1) or to determine general acceptance of such technology (Section 2.2). While there have been additional studies that have examined less related issues, for example, how various demographic or health factors affect technology use and acceptance (e.g., CREATE [2015], Crabb and Hanson [2014], and Rosen et al. [2013]), this section focuses on deaf participants evaluating sign language animations.

2.1. Demographics in Prior Studies

From the perspective of considering how the demographic characteristics or the prior technology experience of participants in a study may affect the results collected, we have examined prior sign language animation studies to note the types of participant characteristics or technology experience/attitudes that researchers reported. Our goal

Table I. Demographic and Technology-Experience Characteristics Reported in Example User Studies

Paper	*Demographic*	*Technology*	*Attitudes*
Hayward et al. [2010]	age, gender, describe, profession	computer expertise	animation usage
Kennaway et al. [2007]	age, gender, describe, signing frequency, preferred language		animation usage
Verlinden et al. [2001]	age, gender, preferred language		attitude to avatar
Gibet et al. [2011]	age, gender, describe, self-reported sign language skills, location		attitude to avatar

was to understand the diversity of participants in prior studies and the types of data that researchers commonly collect. While there have been a small number of studies that only published minimal information, for example, the number of participants and how they are self-identified as deaf or hard-of-hearing [Moemedi 2010; Yang et al. 2014], in general, the trend in the field is to include more information about the sampled population. Table I lists some examples of representative papers in the field, and similar patterns may be found when examining larger surveys of prior evaluation studies (e.g., Ebling and Glauert [2015], Huenerfauth and Kacorri [2014], Huenerfauth et al. [2008], Kacorri and Huenerfauth [2014], and Kipp et al. [2011]).

It is common for researchers to report the age range of participants, the gender ratio of participants, and the ratio of participants identifying as deaf/Deaf or hard-of-hearing (this characteristic is labeled by the term “describe” in Table I, since it refers to how individuals describe themselves). There is wider variation in how studies measure and report the level of sign language skill of their participants: Some use the concept of how often people use signing (e.g., “signing frequency” in Table I) or the individual’s own characterization of their ability (e.g., “self-reported sign language skills”). It is much less common for researchers to report information about their participants’ usage and experience with technology: for example, Hayward et al. [2010] included questions about “computer expertise,” but the other listed papers did not. While they did not report any specific information about their participants, researchers in Verlinden et al. [2001] commented that only those participants who were unfamiliar using the Internet had negative attitudes towards their avatar; they stated, “*This suggests that acceptance of the avatar is greater for web-surfers and that this acceptance may increase as a person becomes more familiar with the Internet.*” Aside from their skill in technology, an individual’s attitude about it may be relevant to consider, especially when a study asks participants to give subjective ratings about sign language animations. Here, there is further variation in whether researchers asked participants about their attitude towards animated avatars (“attitude to avatar”) or their views about the future potential of signing animations in different real-world contexts (“animation usage”). In a few cases, researchers have asked participants to suggest where they could imagine this technology being applied, for example, as an educational tool [Hayward et al. 2010] or for disseminating information in public spaces [Kennaway et al. 2007].

Whereas Table I listed some of the characteristics reported in some prior studies, Table II lists the range of values for these characteristics, for the same set of papers as those in Table I. There is wide variation in the demographic characteristics of the individual participants in prior sign language animation evaluation studies. For example, there is especially wide variation in how researchers assess the signing skills of participants to determine whether they have sufficient fluency or native-level skill to participate in the study; for example, some described what language their participants preferred [Kennaway et al. 2007; Verlinden et al. 2001] and others described how often they used signing [Kennaway et al. 2007].

A key question arises from examining this table: *Do these differences in the demographic characteristics of the population of users in the study have an impact on the*

Table II. Demographic Profile of Participants in Prior Studies

Paper	*Age Range*	*Female: Male*	*Describe*	*Assessing Signing Skills*
Hayward et al. [2010]	35–50	4:1	Deaf	Deaf educators
Kennaway et al. [2007]	16–66	"slightly less female"	"most were deaf, some were hard-of-hearing"	"all were good signers... all using signing on a daily basis"
Verlinden et al. [2001]	20–53	5:4	deaf	Some had preference for sign language; others had no preference between signing or text.
Gibet et al. [2011]	19–56	18:7	17 deaf, 8 hearing	8 "good," 6 "very good," 11 "native/expert"

comprehension scores or subjective judgments of the participants? If the answer to this question is that it does, and if there is wide variation in the set of participants in prior studies (especially if these characteristics of participants are not measured or reported), then it would be difficult to compare the results across various studies that have evaluated sign language animation technologies. In that case, having more information about the participants in the study would make it easier to compare the results across different studies (so that we would know whether a particular set of participants might have been predisposed to have positive or negative evaluations of sign language animations). Thus, the goal of the two studies presented in this article (the regression study described in Sections 5 and 6 and the subsequent experimental study in described in Section 7) is to identify demographic characteristics or technology-experience/attitude factors that relate to user's scores in evaluation studies. Based on these results, we will propose a set of standard questions that could be asked of participants in a user study to evaluate this technology (with the goal of encouraging future researchers to gather and report these characteristics about their participants in publications) to facilitate comparison of results across papers.

There is good reason to think that the answer to the question in the previous paragraph may indeed be "yes": Some prior studies have included anecdotal evidence of relationships between (a) certain participant characteristics and (b) the subjective judgments or comprehension scores for sign language animation (e.g., the "web-surfers" comment in Verlinden et al. [2001]). However, due to the relatively small sample size of most prior studies, researchers rarely present quantitative results for subpopulations. We are not aware of any prior study that conducted an exploration of whether a large variety of participant characteristics may relate to evaluation scores for sign language animation.

2.2. Acceptance of Multiple Signing Avatars

In order to make the results of our study as generalizable as possible for the field of sign language animation technology, Section 4 will discuss how we have included animations with virtual human avatars produced using a variety of modern sign language animation platforms (so that the results are not specific to a particular platform). Since Kipp et al. [2011] carried out the most comprehensive study to date with participants evaluating multiple sign language avatars, in this section, we position our research in relation to this most-closely related prior work.

In a focus-group study, eight native signers of German Sign Language were presented with six avatars signing content in different sign languages, and they commented on their quality [Kipp et al. 2011]. In fact, participants viewed some animations in American Sign Language and other languages that were unfamiliar to them; in contrast, in our study described in Section 5, participants were shown animations in a language in which they were fluent (ASL). Further, researchers in Kipp et al. [2011] showed participants some hand-animated avatars (produced through a painstaking process of carefully posing the character). While the resulting hand-produced animations can be quite beautiful, they are time-consuming to produce and do not address the maintenance efficiency issue discussed in Section 1. Current sign language animation research focuses on *synthesized* animation, in which software automatically selects aspects of the movement to allow for generation of animations from a sparse input script. Section 4 describes how our new study utilized stimuli containing human avatar animation that was synthesized (not hand-animated).

Kipp et al. [2011] also conducted an online survey (N = 317), in which participants rated three avatars (one was hand-animated) on a 5-point scale in regard to comprehensibility, facial expression, naturalness, charisma, movements, mouthing, appearance, hand shapes, and clothing. The hand-animated avatar received higher scores. In our new study, in addition to subjective ratings, we include objective comprehension questions to measure participants' understanding because prior research has demonstrated that self-reports of understanding typically have low correlation to a participant's accuracy at answering comprehension questions [Huenerfauth et al. 2008].

Notably, in both the focus group and the online survey, the authors observed higher scores in response to the questions "Do you think avatars are useful?" and "Do you think Deaf people would use avatars?" when asked at the end of the study (compared to the beginning). The authors speculate that additional exposure to animations influenced participants' responses. To investigate this issue, in our new study, we include a question about whether participants had previously seen computer animations of sign language (details in Section 5.2).

Participants in Kipp et al. [2011] also suggested use-cases for signing avatars, including public transit, movies/entertainment, government and educational websites, and other areas. In our new study, we also asked participants to judge the usefulness of signing avatars in various contexts: information on websites, for public places (e.g., airport, train station), as a virtual interpreter in a face-to-face meeting, as a virtual interpreter for telephone relay, etc. Section 5.2 will summarize the responses of participants to these questions.

While Kipp et al. [2011] collected some demographics (gender, age, deaf/hard-of-hearing/hearing, and profession), those researchers were not focused on the primary research question of this article (the connection between participant characteristics and their responses). For that reason, they did not analyze the data to look for relationships between these factors and the survey responses. The study presented in Section 5 includes a regression analysis to identify demographic and experience factors related to the participants' subjective responses and comprehension scores.

Given the online modality of the study presented in Kipp et al. [2011], there is a possibility that participants could have been more comfortable using the Internet than the general population. In our new study (Sections 5 and 6), we conduct an in-person survey in which participants evaluate sign language animations; members of our research team traveled to meet participants at convenient locations. Our goal was to encourage the participation of less technology-savvy individuals and to enable us to confirm that participants met our study criteria (and that they were accurately reporting their demographic data, at least for those characteristics apparent to the researcher).

3. SELECTING QUESTIONS TO COLLECT INDEPENDENT VARIABLES

With a goal of examining whether metrics relating to participants' demographics (e.g., age, gender) or technology experience/attitudes can explain some of the subjective-judgment and comprehension-question scores collected in experiments to measure the quality of sign language animation systems, this article describes a survey and regression analysis (Sections 5 and 6) and a subsequent experimental study (Section 7). This section explains the design of our questionnaire for recording the independent variables about participants' characteristics, which were used in our multiple-regression models in Section 6. This section will also explain the origin of any questions that were adapted from survey instruments that were presented in prior work of other authors (e.g., Rosen et al. [2013]). A subset of these questions was used in the subsequent experimental study described in Section 7.

While some researchers have explored the design of fully online surveys of deaf users containing both ASL and English (e.g., Tran et al. [2010]), we chose to conduct our survey in-person, with a human signer asking questions in ASL on a laptop screen and a paper answer sheet (with questions redundantly appearing in English, to aid the participant in aligning the video and paper). Given that our study included hard-of-hearing participants, the inclusion of English was considered important, and given our aim to include older participants in the study, a "low tech" paper answer sheet was preferable. Many questions were adapted from preexisting English surveys (Section 5.2); so a professional ASL interpreter (bachelor's degree in interpreting and master's in information technology) translated items into ASL. Deaf members of the research team checked that subtleties of meaning were preserved. Several takes of each question were recorded so that we could select the best version for the questionnaire. Example videos appear in the online appendix to this article in the ACM Digital Library.

3.1. Demographic Questions

Demographic questions were selected by assembling items that were asked in prior experimental studies (e.g., Huenerfauth and Kacorri [2015]), and questions asked in studies surveyed in Section 2. Next, the demographic questions are listed, preceded by the "codename" of the response variables used in our regression models in Section 6.

Gender: What is your gender? (male, female, other)
Age: How old are you? (Note: After collecting data from participants, as described in Section 5, we noticed a gap in the age range 35–42, so instead of treating age as a continuous variable, we binned it into three groups: 18 to 24, 25 to 34, and 43 to 59, and we relabeled the variable as **AgeGroup**.)
Describe: How do you describe yourself? (deaf/Deaf, hard-of-hearing, hearing, other)
WhenBecome: At what age did you become deaf or hard-of-hearing? (Note: No hearing participants were in this study.)
WhenLearn: At what age did you begin to learn ASL? (Note: all participants in this study were ASL signers.)
ParentsAre: Are your parents deaf/Deaf? (yes, no)
ParentsUse: Did your parents use ASL at home? (yes, no)
SchoolType: What type of school did you attend as a child? (a residential school for deaf students, a daytime school for deaf students, or a mainstream school)
SchoolASL: Did you use ASL at this school? (yes, no)
Education: Which describes your current level of education? (I did not graduate high school, I graduated high school, I graduated college, I have a bachelor's degree, I have a graduate degree)
HomeASL: Do you use ASL at home? (yes, no)
HomeEnglish: Do you use English at home? (yes, no)
WorkASL: Do you use ASL at work? (yes, no)
WorkEnglish: Do you use English at work/school? (yes, no)

3.2. Technology Experience and Attitudes

In order to measure participants' frequency of technology use, the InternetSearch and MediaSharing subscales were used from the *Media and Technology Usage and Attitudes Scale* [Rosen et al. 2013]; scoring is based on the participant's response (e.g., Never, Monthly, Weekly, Once a day, etc.) to how frequently they engaged in various activities (listed next) on computers, laptops, tablets, or mobile phones:

InternetSearch: How often do you search the Internet for news? How often do you search the Internet for information? How often do you search the Internet for videos? How often do you search the Internet for images or photos?

MediaSharing: How often do you watch TV shows, movies, etc., on a computer, laptop, tablet, or smartphone? How often do you watch video clips on a computer, laptop, tablet, or smartphone? How often do you download media files from other people on a computer, laptop, tablet, or smartphone? How often do you share your own media files on a computer, laptop, tablet, or smartphone?

Using the same scoring, we created an ASLChat subscale:

ASLChat: How often do you have a signing (ASL) conversation with someone using a video phone? How often do you have a signing (ASL) conversation with someone using a computer, laptop, tablet, smartphone?

We asked participants to indicate how often they played video games (and thereby may have more experience viewing animated humans) by selecting one of three frequency ranges (below), which we coded as "advanced," "intermediate," and "beginner."

GameGroup: How often do you play games on a computer, game console, or phone? (several times a day, between once a day and once a week, less than once a week)

Next, participants were asked about their perceptions of the benefits of technology, using the PositiveAttitudes subscale of Rosen et al. [2013], in which the score is the average of responses to individual statements listed below (Strongly agree = 5, Agree = 4, Neither agree nor disagree = 3, Disagree = 2, Strongly disagree = 1):

PositiveAttitudes: It is important to be able to find any information whenever I want to online. It is important to be able to access the Internet any time I want. It is important to keep up with the latest trends in technology. Technology will provide solutions to many of our problems. With technology anything is possible. I accomplish more because of technology.

Participants' impression of computer complexity was measured using two Computer Questionnaire questions from the October 2014 PRISM survey [CREATE 2015], using identical Likert scoring as earlier.

ComputerComplex: Computers are complicated. Computers make me nervous.

Finally, at the end of the questionnaire, users were asked to indicate their agreement with a series of statements (below) to evaluate their overall attitude of the usefulness of ASL animations in a variety of contexts; this novel set of Likert-type items was inspired by questions in Gibet et al. [2011], Kennaway et al. [2007], and Kipp et al. [2011]. Finally, users were also asked if they had previously seen computer animations of ASL:

AnimationAttitude: Computer animations of sign language could be used to give information on a website. Computer animations of sign language could be used to

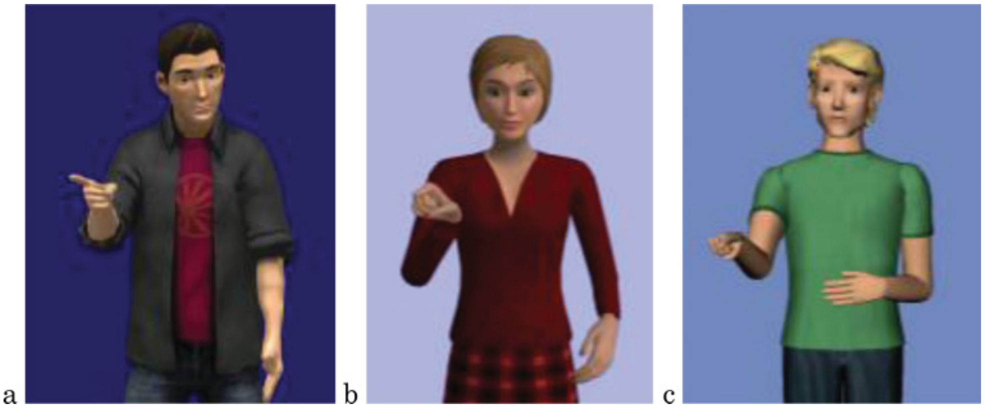

Fig. 1. Screenshots from the three avatars shown in the study: (a) EMBR, (b) JASigning, (c) VCom3D.

give information in a public place (e.g., airport, train station). Computer animations of sign language could be used as an interpreter in a face-to-face meeting. Computer animations of sign language could be used as an interpreter for a telephone relay. I would enjoy using computer animations of sign language. Other people would enjoy using computer animations of sign language.

SeenBefore: Before today, had you ever seen a computer animation of sign language? (yes, no)

4. SELECTING STIMULI AND QUESTIONS TO COLLECT DEPENDENT VARIABLES

In Section 2.2, we discussed how prior researchers in Kipp et al. [2011] displayed animations of multiple sign languages and animations that were hand-animated; we explained why we decided to display only **synthesized** animations of **ASL** in our current study. However, we wanted the results of our study to be generalizable to a variety of ASL signing avatars, with different appearance, rendering technologies, automation capabilities, and motion synthesis. Thus, we decided to display animations of three avatars synthesized by different state-of-the-art animation platforms [Jennings et al. 2010; Kacorri and Huenerfauth 2015; VCom3D 2015]. While in the prior study of Kipp et al. [2011], each avatar performed a different message, to control for this in our study, we selected three short ASL stories from a stimuli and comprehension-question collection made available to the research community in Huenerfauth and Kacorri [2014]. Specifically, we selected three stimuli (codenames N2, W2, and Y3) that had been rated as being the most understandable in an earlier study by Kacorri and Huenerfauth [2015]. Example stimuli from the current study appear in the online appendix to this article in the ACM Digital Library.

—***EMBR:*** The open source EMBR platform [Heloir et al. 2011], extended with ASL handshapes and detailed upper-face controls using the MPEG-4 Facial Animation standard [ISO 2004], was used to produce the first type of stimuli, as shown in Figure 1(a). A team of native ASL signers selected key poses to define each sign in the system's lexicon, in order to create the avatar's hand movements. To produce the face and head movements, video recordings of a native ASL signer performing the stimulus were analyzed by the Visage Face Tracker, an automatic face tracking software that provides MPEG-4 compatible output, as described in Kacorri [2016]. After extracting the facial features and head pose from the video of the human signer, this data was used to automatically drive the animated character by converting the face and head movement information into the script language supported by the EMBR platform, as described in Kacorri and Huenerfauth [2014].

—***JASigning:*** Next, the free Java Avatar Signing (JASigning) system [Jennings et al. 2010] was used to produce the second type of stimuli, as shown in Figure 1(b). To produce the movements of the character, all of the individual ASL signs were notated in the Hamburg Notation System (HamNoSys) [Prillwitz et al. 1989] by a deaf researcher who consulted video recordings of an ASL native signer performing each stimulus. The HamNoSys notation system, which serves as the input for the JASigning platform, has approximately 200 symbols that can symbolically specify handshape, hand position, location, and movement. Information about the nonmanual components (e.g., eyebrow movement, eye gaze, and head movement) is included in the SiGML code [Hanke 2001], an XML representation for HamNoSys, but time-alignment of nonmanuals with the manual signs requires careful adjustment (e.g., Ebling and Glauert [2015]).

—***VCOM:*** Finally, a commercially available ASL authoring tool, VCom3D Sign Smith Studio [VCom3D 2015], was used to produce the third type of stimuli, as shown in Figure 1(c). This software enables users to produce animated ASL sentences by arranging a timeline of signs from a prebuilt or user-defined vocabulary. It includes a library of facial expressions that can be applied over a single sign or multiple manual signs. Both the hand movements and facial expressions of the avatar for the three stimuli were created by native ASL signers at a key-pose level. In addition, both the VCOM and EMBR animations shared similar hand movements.

To collect responses from participants about their comprehension of the animation, we used a set of objective questions: After viewing each of the animations, an on-screen video of a native ASL signer asked participants four fact-based comprehension questions about the information conveyed in the animation. Participants responded to each question on a 7-point scale from “definitely no” to “definitely yes.” As described in Huenerfauth and Kacorri [2014], a single “Comprehension” score for each animation can be calculated by averaging the scores of the four questions.

Next, the participants were asked to respond to a set of questions that measured their subjective impression of the animation, using a 1-to-10 scalar response. Each question was conveyed using ASL through an onscreen video, and the following English question text was shown on the questionnaire:

(a) Good ASL grammar? (10 = Perfect, 1 = Bad)
(b) Easy to understand? (10 = Clear, 1 = Confusing)
(c) Natural? (10 = Moves like person, 1 = Like robot)
(d) Was the signer friendly? (10 = Friendly, 1 = Not)
(e) Did you like the signer? (10 = Love it, 1 = Hate it)
(f) Was the signer realistic? (10 = Realistic, 1 = Not)

Questions (a)–(c) have been used in many prior experimental studies and were included in the collection of standard stimuli and questions that was released to the research community by Huenerfauth and Kacorri [2014]. Questions (d)–(f) were inspired by Kipp et al. [2011]. To calculate a single “Subjective” score for each animation, the scalar-response scores for the six questions were averaged.

5. DATA COLLECTION FOR STUDY #1 (SURVEY DATA FOR REGRESSION ANALYSIS)

This article describes two sets of studies with deaf participants evaluating animations: Study #1 was a survey conducted with 62 participants whose responses were later analyzed using a regression-based analysis. Study #2 was an experimental study with 57 participants that was designed to empirically evaluate hypotheses suggested by the regression analysis in Study #1. This section describes the data-collection process used for Study #1, followed by a summary of the characteristics of the participants. Section 6 will present the regression analysis of these data, and Section 7 will describe the subsequent experimental Study #2.

Our laboratory has over a decade of prior experience in conducting empirical evaluations of sign language animation with deaf participants. For example, in Huenerfauth and Kacorri [2015], we investigated key methodological considerations in conducting a study to measure comprehension of sign language animations with deaf users, including the use of appropriate baselines for comparison and the appropriate method for presenting comprehension questions and instructions. In Huenerfauth et al. [2008], we describe the advantages of having deaf researchers conduct experimental studies in ASL.

In Study #1, a deaf researcher (coauthor) and two deaf undergraduate students (native ASL signers) recruited and collected data from participants, during meetings conducted in ASL. Potential participants were asked if they had grown up using ASL at home or had attended an ASL-based school as a young child. Initial advertisements were sent to local email distribution lists and Facebook groups. Our study (N = 62) was completed during a 4-week data-collection period, a short time frame made possible due to the many people who are deaf and hard-of-hearing associated with Rochester Institute of Technology (RIT) or living in Rochester, NY. Given the use of a university campus as a basis for some of the recruiting, we found it easier to identify younger participants (especially college-aged students); the process of recruiting older participants took additional time and effort. The research team used personal contacts in the Deaf community to identify participants, especially older adults, who were less likely to be recruited through electronic methods. The advertisement included contact information for a deaf researcher, including an email address, videophone, and text messaging (mobile phone). Research team members also attended local Deaf community events (e.g., the Deaf Club) to advertise the study.

Researchers met participants around Rochester to conduct the 70-minute survey, using a laptop with video questions in ASL. After participants answered the demographic and technology-experience questions, they viewed a sample animation, to become familiar with the experiment setup and the questions they would be asked about each animation. (This sample animation used a different avatar than the other animations shown in the study.) As described in Section 4, after viewing each of the animations, participants answered comprehension questions about the animation's content and subjective questions about their opinion of the animation.

Before presenting a regression model that investigates the relationship between our dependent and our independent variables, in the following two subsections, we will first briefly summarize the characteristics of our participants (e.g., how many people fell into each independent variable category). The visualizations that appear in Figures 2–10 were produced using "likert" package of R [Speerschneider and Bryer 2013].

5.1. Demographic Characteristics

Of the 62 participants recruited for the study, 43 participants learned ASL prior to age 5, 16 had been using ASL for over 9 years, and the remaining three learned ASL as adolescents, attended a university with classroom instruction in ASL, and used ASL daily to communicate with a significant other or family member. There were 39 men and 23 women of ages 18–59 (mean 25.73, standard deviation 10.47). Among those participants over age 43 (average age 53.14), there were four men and two women who learned ASL prior to age 9, five self-reported to be deaf/Deaf and one hard-of-hearing.

The male-to-female ratio of our participant pool was similar to that of some prior surveys of the U.S. population, which indicate higher rates of hearing impairment

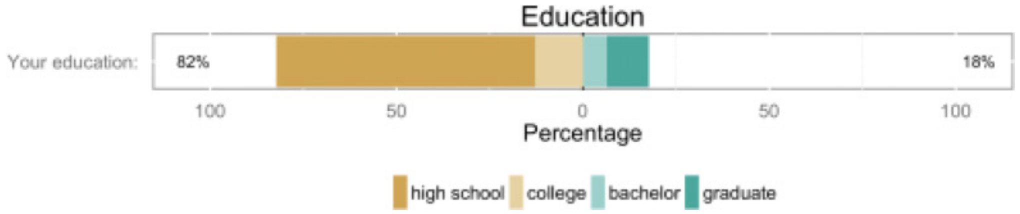

Fig. 2. Responses to education question in our study of 62 participants.

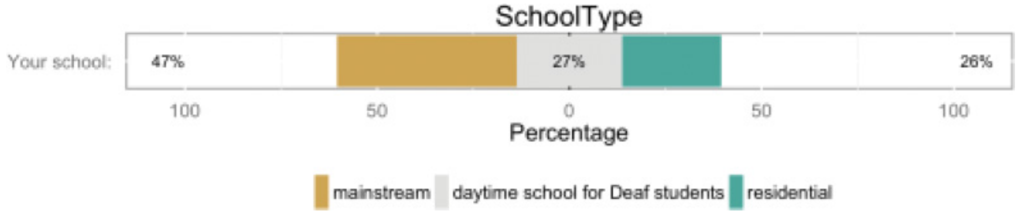

Fig. 3. Responses to the question about whether the participant attended a mainstream school, a daytime school for deaf students, or a residential school for deaf students as a child.

among males (e.g., the 1990–1991 National Health Interview Survey, as reported in Holt et al. [1994]).

Compared to prior surveys of the U.S. population (e.g., as discussed in Mitchell et al. [2006]), our participant pool was relatively young, which is not surprising given how some of our recruitment was on a university campus. However, a further distinction is that we were focused on recruiting individuals who were fluent ASL signers, which is a different demographic group than those who are counted in some national surveys in the U.S. that include many people who became deaf or hard-of-hearing later in life, whom are less likely to become fluent ASL signers. For instance, our participants had WhenBecome scores ranging from 0 to 14 (mean 1.5, standard deviation 2.83), and WhenLearn scores ranging from 0 to 19 (mean 5.08, standard deviation 4.63).

Figure 2 summarizes our participants' responses to the question about their educational background. Based on U.S. national surveys, it has been reported that among the "severely to profoundly hearing-impaired population," 44% had not graduated high school, 46% had a high school diploma, 5% had graduated college, and 5% had a postgraduate degree [Blanchfield et al. 2001]. Our participants had relatively higher levels of educational attainment than these previously reported national averages: None of the participants in our study indicated that they had not graduated high school, 69% had a high school diploma, 13% had completed some college, 7% had a bachelor's degree, and 11% had a postgraduate degree.

Figure 3 presents the responses of our participants to the question about the type of school participants attended as a child, that is, whether it was a mainstream school (47% of our respondents), a daytime school for deaf students (27% of respondents), or a residential school for deaf students (26% of respondents). These responses were relatively similar to those reported in Feldman et al. [2000] on data from the Gallaudet Annual Survey in 1997–1998, which indicated that children with "profound hearing impairment" ages 3–17 were in a mainstream school (55%), residential school for deaf students (32%), and daytime school for deaf students (13%).

Figure 4 collects all of the participants' responses to the polar (yes-or-no) questions on the demographic portion of our questionnaire. In our study, 21% of participants indicated that their parents were deaf: Surveys of the U.S. population have indicated that more than 90% of deaf children are born to hearing parents, and recent analyses

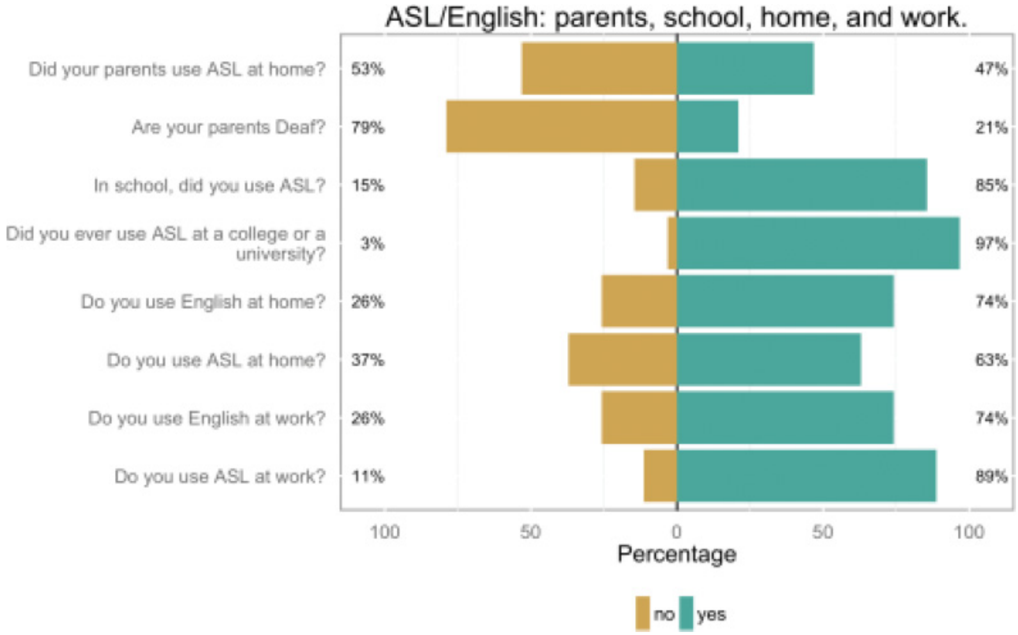

Fig. 4. Responses to demographic survey questions about usage of ASL at home during childhood and contemporaneous use in work, school, or at home.

have suggested that this percentage may be even higher [Karchmer and Mitchell 2004]. The higher percentage of participants reporting deaf parents in our study may be due to our recruiting fluent ASL signers, who may have been more likely to use ASL at home. We speculate that our recruiting of fluent signers (along with our recruitment on a university campus with ASL use and among a local community in Rochester, NY, of ASL signers) may have also led to the relatively high responses for the questions about usage of ASL at home during childhood and the use of ASL currently in work or school. The high use of ASL among participants in this study is reasonable given that we are interested in evaluating sign language animation technology, which may be specifically targeted for use among individuals who use sign language in their daily life.

5.2. Technology-Experience Characteristics

Participants in Study #1 also responded to questions about their technology experience and their attitudes about technology and computer animation. Figures 5–10 summarize these responses. In general, the participants in the study had largely positive responses to the question items on the PositiveAttitudes subscale; these responses may partially be explained by the relatively young demographic.

Figure 6 presents responses to the questions about participants' perception of the complexity of computers and the degree to which computers make them nervous. In general, participants disagreed with these statements, indicating a level of comfort with computing technology.

Figure 7 summarizes participants' responses to the novel scalar-response questions presented in this study that measured frequency of use of technologies for remote video conferencing using sign language. Very few participants in the study reported that they had never used such technology, indicating its increasing popularity. In this visualization, we chose to align the bars for each response so that the boundary between

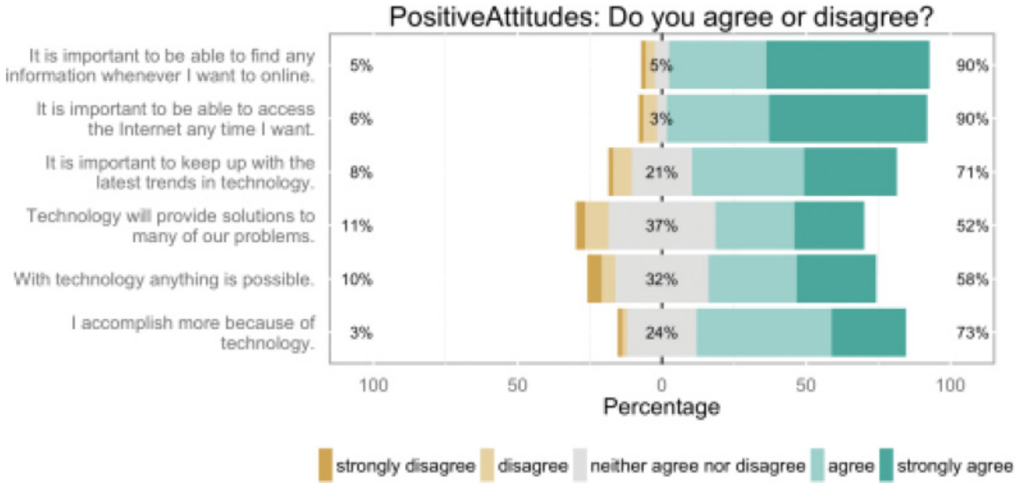

Fig. 5. Responses to questions about participants' attitudes about technology.

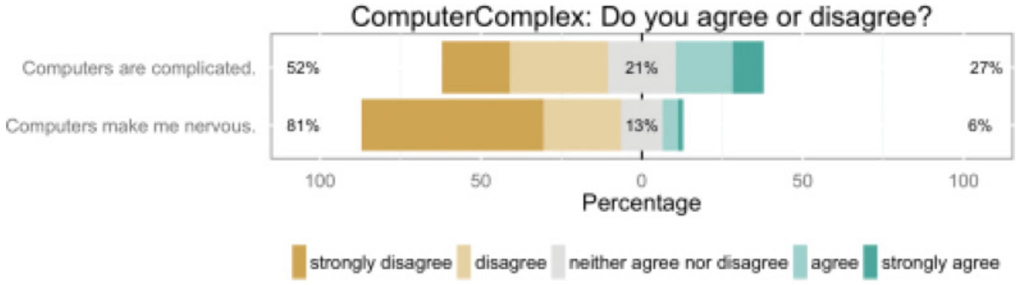

Fig. 6. Responses to questions about perception of computer complexity.

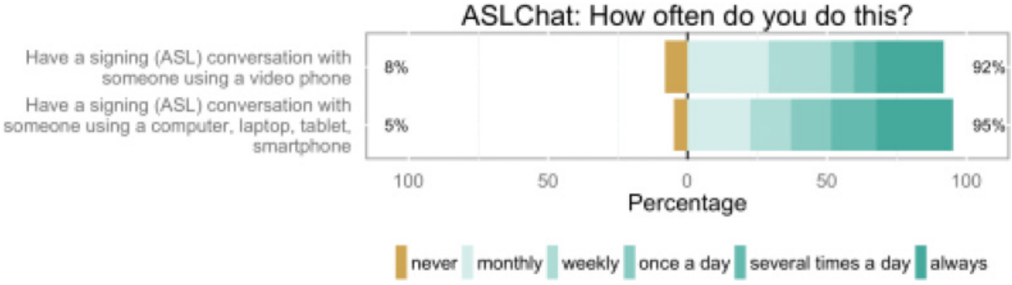

Fig. 7. Responses to questions about frequency of use of technologies for remote video conferencing using sign language.

the "Never" and "Monthly" responses are along the midline of the graph. The rationale for our choice is that those respondents who selected "Never" to these items indicate "zero" usage of that technology.

Figures 8 and 9 present the responses of participants to the questions that were on the InternetSearch and MediaSharing subscales of the *Media and Technology Usage and Attitudes Scale* [Rosen et al. 2013]. We observed responses indicating very frequent use of the Internet for searching information, which is not surprising given the ubiquity of this technology. Responses for the MediaSharing questions were relatively

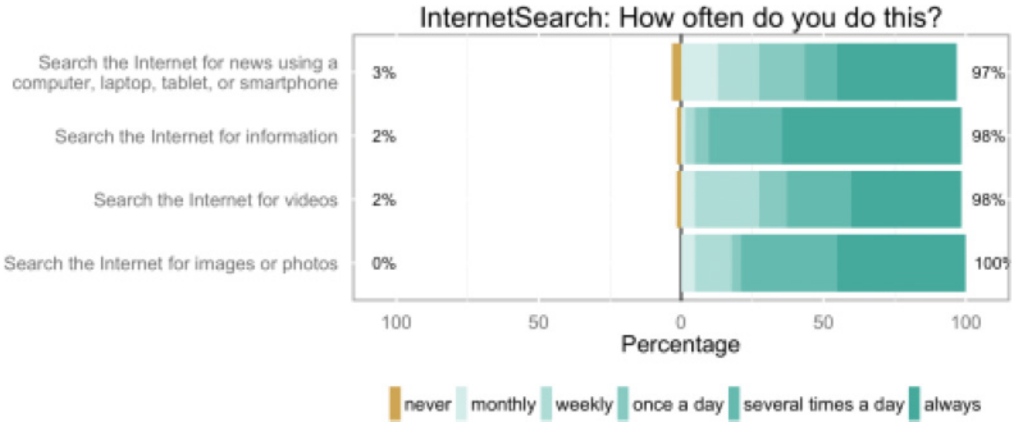

Fig. 8. Responses to questions about the use of the Internet for searching.

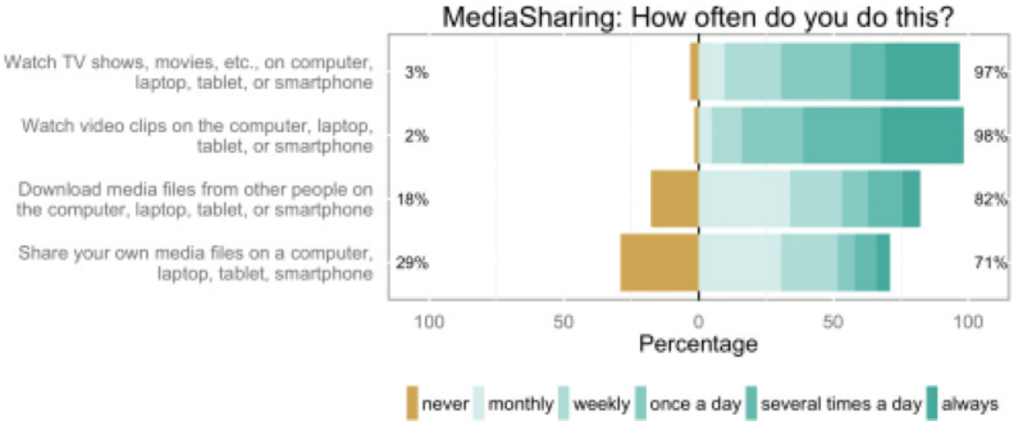

Fig. 9. Responses to questions about the use of electronic devices for viewing and sharing media and videos.

lower, especially for those questions asking about downloading media from others or individuals posting their own media online.

Figure 10 presents the results of the novel Likert-type question items on our questionnaire that were designed to measure participants' attitudes about computer animation technology being used for sign language. Compared to the previously presented questions about attitudes and towards technology in general (e.g., Figure 5), we found relatively lower scores for computer animations of sign language. Among the questions that asked about various contexts of use, participants were especially skeptical about the use of this technology to provide interpretation in face-to-face meetings or during telephone relay conversations. Participants expressed more agreement with statements about the use of this technology on websites or to provide information in public places. A relatively wide diversity of responses was observed in response to the question about whether the individual would enjoy using ASL animations, with participants answering somewhat more positively as to whether "other people" would enjoy using ASL animation technology.

Based on the participants' responses to the question about their frequency of playing video games, we found that respondents tended to select the low-frequency and high-frequency extremes of the response scale. Specifically, 33 respondents said that they

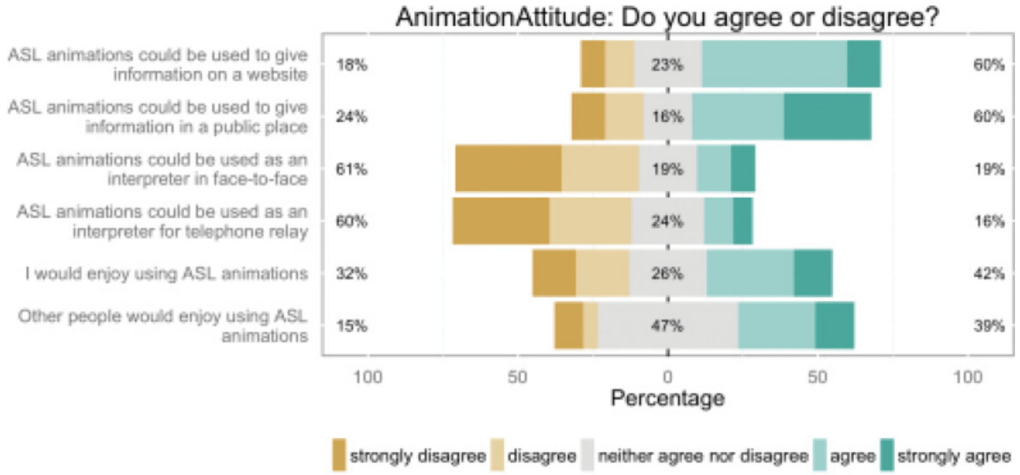

Fig. 10. Responses to questions about participants' attitudes about the use of sign language computer animation technology.

played less than once a week (we labeled them as "beginners"), 9 said that they played between once a day and once a week (we labeled them as "intermediate"), and 20 said that they played several times a day (we labeled them as "advanced").

Participants in the study were also asked a polar (yes-or-no) question about whether they have ever seen computer animations of sign language prior to the current study. Among our 62 participants, 29 said that they had previously seen computer animations of signing, and 33 indicated that they had not previous seen such animations.

6. MULTIPLE-REGRESSION ANALYSIS AND RESULTS

This section presents a multiple-regression analysis of the data collected during Study #1, in order for us to examine whether demographic factors relate to participants' responses to subjective and comprehension questions about ASL animations. In addition, this analysis examined whether variance in scores could be explained by participants' technology experience and attitudes.

For the multiple-regression models discussed in this section, our independent variables included all of the "Demographic" and "Technology" metrics, listed in Section 3. Our dependent variables included the "Comprehension" and "Subjective" scores described in Section 4. To facilitate easier comparison among coefficients of scalar and binary predictors, many researchers (e.g., Crabb and Hanson [2014]) follow the recommendation of Gelman [2008] that continuous-value variables be normalized by dividing the individual participant metrics by two times the group standard deviation. We have also followed this procedure for all of the continuous independent variables in this study.

We actually trained two separate models for each of our dependent variables (Subjective and Comprehension): Model 1 was based upon Demographic variables only, and Model 2 was based upon both Demographic and Technology variables. The rationale for this choice is that while some prior authors have reported limited Demographic data about the participants in their studies, the set of Technology questions presented in this article is novel. Since we had recorded many Demographic and Technology variables (Section 3), it was important to explore combinations of variables in a systematic manner.

Table III. Multiple-Regression Model—Comprehension
Significance codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

	Estimate	*Std. Error*	*t score*
Model 1: Demographic	Model 1: Adj. $R^2 = 0.256$ ($p<0.005$)		
AgeGroup (25,34)	−0.344	0.195	−1.768 .
AgeGroup (35,)	−0.094	0.207	−0.452
Describehard-of-hearing	−0.242	0.149	−1.629
WhenBecome	0.204	0.126	1.624
WhenLearn	0.164	0.152	1.081
ParentsAreyes	0.252	0.166	1.516
SchoolASLyes	0.336	0.183	1.838 .
HomeASLyes	−0.177	0.147	−1.204
WorkEnglishyes	0.292	0.152	1.923 .
SchoolTypeMainstream	−0.092	0.146	−0.630
SchoolTypeResidential	0.575	0.169	3.407 **
Model 2: Demogr. & Tech.	Model 2: Adj. $R^2 = 0.382$ ($p<0.0001$)		
Gendermale	0.273	0.126	2.168 *
Describehard-of-hearing	−0.317	0.135	−2.338 *
WhenBecome	0.217	0.117	1.857 .
HomeASLyes	−0.207	0.125	−1.655
SchoolTypeMainstream	−0.029	0.140	−0.208
SchoolTypeResidential	0.662	0.151	4.380 ***
InternetSearch	−0.493	0.140	−3.513 ***
PositiveAttitudes	0.249	0.118	2.105 *
ASLChat	0.181	0.129	1.402
GameGroupBeginner	−0.307	0.129	−2.377 *
GameGroupIntermediate	−0.283	0.202	−1.399
SeenBeforeyes	0.162	0.119	1.355

To build models of all possible subsets of features (to identify the model with the highest adjusted R-squared value), we used the "leaps" package [Lumley and Miller 2009]. (The R-squared metric indicates the total variability accounted for by the model.) For Model 1, the input to "leaps" was all Demographic variables only. For Model 2, the input to "leaps" was all Demographic and all Technology variables. For all models, we evaluated the collinearity of the independent variables (that were selected by "leaps") by verifying that their variance-inflation was less than 2 [Fox and Monette 1992].

Table III summarizes the models built during the regression analysis for the Comprehension dependent variable.[2] In Model 1 (using demographic variables only as independent variables), the type of school that the participant attended had the largest coefficient (see the values in the "Estimate" column): attending a residential school for deaf students had a positive relationship with the participant's success at answering comprehension questions.

In Table III, Model 2 contained both demographic and technology variables as independent variables, and a relationship between SchoolType and Comprehension is still present. Gender, Describe, InternetSearch, PositiveAttitudes, and GameGroup were also key components of Model 2. This suggests that when considering the results of studies that evaluate participants' comprehension of synthesized ASL animations, some variance in participants' scores can be explained by the demographic and technology characteristics of each participant, for example, their use of the Internet, positive

[2] The *Estimate* column reports the regression coefficient for the variable (how output varies per unit change in variable), *Std. Error* indicates average model error in the variable units (smaller values indicate that the observations are closer to the fitted line), and *t score* is the test statistic used to calculate the p-value for significance testing.

Table IV. Multiple-Regression Model—Subjective
Significance codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1

	Estimate	*Std. Error*	*t score*
Model 1: Demographic	**Model 1:** Adj. $R^2 = 0.153$ ($p< 0.02$)		
Gendermale	– 0.527	0.501	– 1.05
Describehard-of-hearing	0.652	0.576	1.13
WhenLearn	– 0.834	0.542	– 1.54
HomeASLyes	– 1.557	0.591	– 2.63 *
SchoolTypeMainstream	0.659	0.584	1.13
SchoolTypeResidential	– 0.538	0.643	– 0.84
Model 2: Demogr. & Tech	**Model 2:** Adj. $R^2 = 0.335$ ($p<0.0001$)		
WhenLearn	– 0.589	0.486	– 1.21
HomeASLyes	– 1.431	0.499	– 2.87 **
SchoolTypeMainstream	0.685	0.517	1.32
SchoolTypeResidential	– 0.030	0.590	– 0.05
ComputerComplex	0.628	0.426	1.48
MediaSharing	– 1.491	0.448	– 3.33 **
AnimationAttitude	– 1.373	0.448	– 3.07 **

attitude towards technology, and video game exposure. (Section 6.1 includes additional discussion of these factors.)

Table IV summarizes the models built during the regression analysis for the Subjective-scores dependent variable. In Model 1 (using only demographic variables as independent variables), using ASL at home had a significant and downward effect on a participant’s subjective impressions. Using ASL at home was also a significant factor in Model 2, which includes both Demographic and Technology variables. Moreover, AnimationAttitude and MediaSharing were other key components of Model 2. These results suggest that when considering the results of studies that collect subjective judgments about synthesized sign language animations, researchers can expect harsher judgments from participants who use ASL at home, are comfortable with media sharing or downloading, and whose general attitude about sign language animations and their usefulness is not positive.

Figure 11 illustrates how Comprehension Model 2 accounts for significantly more variance than Comprehension Model 1, and the same is true for Subjective Model 2 and Subjective Model 1. An ANOVA was used to compare the models, and p-values are denoted in the graph by *** for $p<0.001$ or by ** for $p<0.01$. Model 2 represented a significant improvement in the amount of Comprehension accounted for between groups from 25.6% to 38.2%. Loosely speaking, this indicates that you can more accurately predict a signer’s success at answering comprehension questions by considering both their demographic characteristics and technology experience/attitudes, rather than relying on their demographic characteristics only. Similarly, there was a significant increase in accounted variance of participants’ subjective impressions of the animations from 15.3% to 33.5%.

6.1 Relative Importance of Features in the Models

Since the goal of our research was to identify which participant characteristics may be most predictive of their response scores (so that we could encourage future researchers to report those characteristics of their participants in publications), we are ultimately interested in a comparison of which of the characteristics were most important in each of the regression models. Henceforth, our discussion will focus only on the best performing models: Comprehension Model 2 and Subjective Model 2, which contained both Demographic and Technology variables.

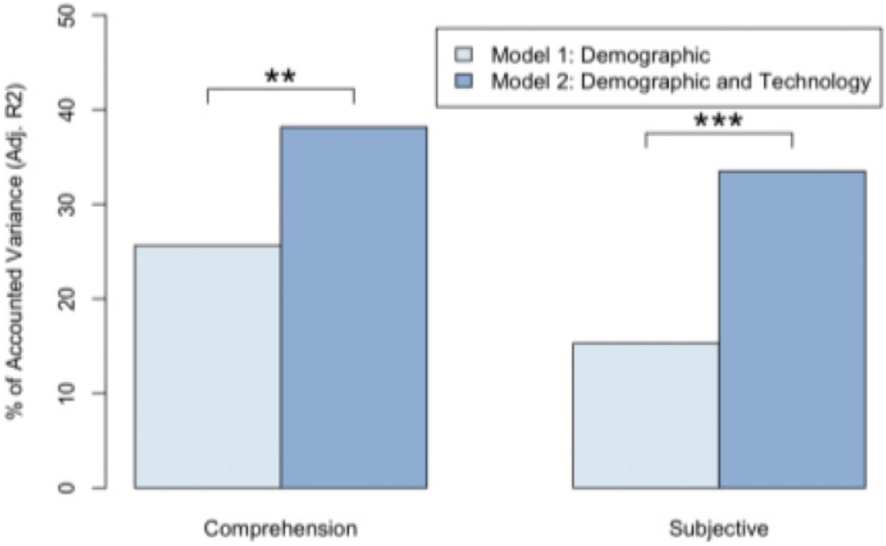

Fig. 11. Regression model comparison summary. (Significance codes: 0 “∗∗∗” 0.001 “∗∗” 0.01.)

In Section 6, we considered each variable’s coefficient (“Estimate” column in Tables III and IV) to roughly identify those with large influence. However, coefficients are sensitive to the “order” in which the variables are considered in the model. For more meaningful interpretation, we calculated the relative importance of each of the variables in Comprehension Model 2 and Subjective Model 2, using the Linderman-Merenda-Gold (LMG) metric [Lindeman et al. 1980], calculated using the “relaimpo” package [Grömping 2006]. This analysis assigns an R-squared percent contribution to each correlated variable *obtained from all possible orderings of the variables in the regression model*. Higher bars in Figure 12 indicate variables with greater importance in the model. We employed bootstrap to estimate the variability of the obtained relative importance value, to determine 95% confidence intervals (shown as whiskers in Figure 12). Importance values may be considered significant when a bar’s whiskers do not cross the zero line in the graph.

For Comprehension Model 2, which contains variables that “leaps” selected through an exhaustive search of all subsets of Demographic and Technology variables, we observe that the variables with highest and significant relative importance were SchoolType, InternetSearch, and GameGroup. Given the much higher relative importance of the SchoolType variable, as compared to the other variables in the model, we focus on this variable in our following discussion:

—***Comprehension and SchoolType***. As discussed in Section 6, attending a residential school seems to have a significant positive relationship with a participant’s comprehension-question scores for synthesized ASL animations. We therefore encourage sign language animation researchers to include this variable in their demographic questionnaire for each study and to report this characteristic of participants in publications. When evaluating the Comprehension scores for their animations, they should consider this factor when comparing their results to those from other studies (whose participant pools may have differed in this characteristic).

—***Comprehension and SeenBefore.*** Another aspect of Figure 12 that may be of interest to sign language animation researchers is the low importance of the SeenBefore variable in this model, which indicates whether the participant had previously seen

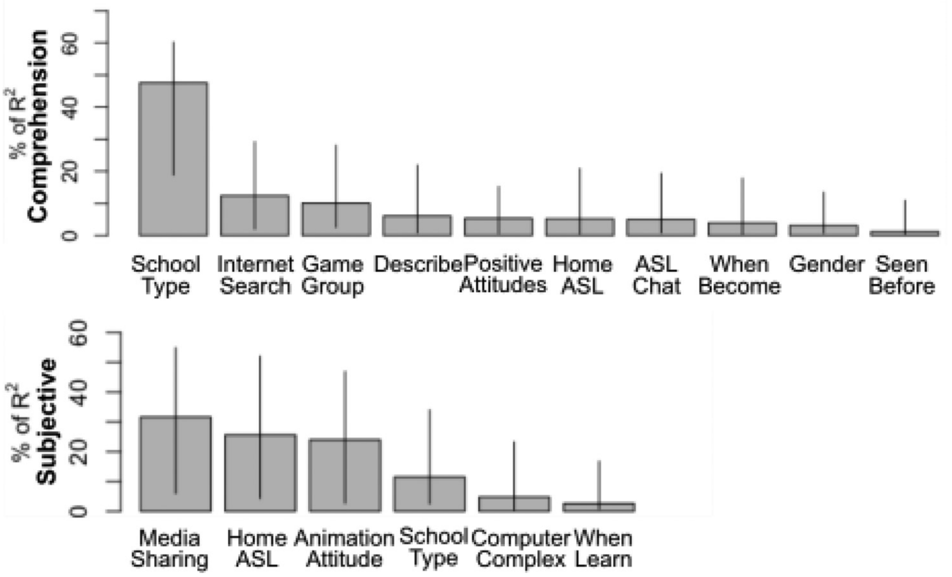

Fig. 12. Relative importance (normalized to sum to 100%) of factors in Comprehension Model 2 and in Subjective Model 2, with 95% bootstrap confidence intervals.

animations of sign language before the study. We noted that prior exposure of a participant to signing avatars did not explain much variance in participants' Comprehension scores. For researchers who conduct user studies with deaf participants to frequently evaluate the progress of their animation software, this finding suggests that participants who have seen prior versions of their animation system may be rerecruited for future studies (with the caveat, of course, that the new study is showing different stimuli). Since there may be a relatively small local Deaf community nearby to some research groups, this is a useful finding. As discussed in Section 5.2, we had a well-balanced sample of participants in this study for the SeenBefore variable (yes = 29, no = 33).

For Subjective Model 2, containing variables that "leaps" selected through an exhaustive search of all subsets of Demographic and Technology variables, we observe that the variables with the highest and significant relative importance are MediaSharing, HomeASL, AnimationAttitude, and SchoolType. While the height of its bar in Figure 12 indicates each variable's importance, the direction of the relationship (positive/negative) is indicated by the sign of the coefficient in the "Estimate" column of Table IV.

—***Subjective and AnimationAttitude.*** We observed a positive relationship between these two variables, which is not a surprising result: If a participant has an overall negative view of the usefulness or likeability of sign language animations *in general* (as measured by the AnimationAttitude scale, Section 3.2), then it is intuitive why they might have lower subjective scores *for a specific animation*.

—***Subjective and MediaSharing.*** Intuitively, we had expected that users with greater technology experience might have higher subjective scores, perhaps due to their possible enthusiasm for technology. On the contrary, we observed that the MediaSharing variable had a negative relationship to participants' subjective scores for animations. We can speculate that users with higher technology experience might have "higher standards" for the acceptable level of quality in an animation.

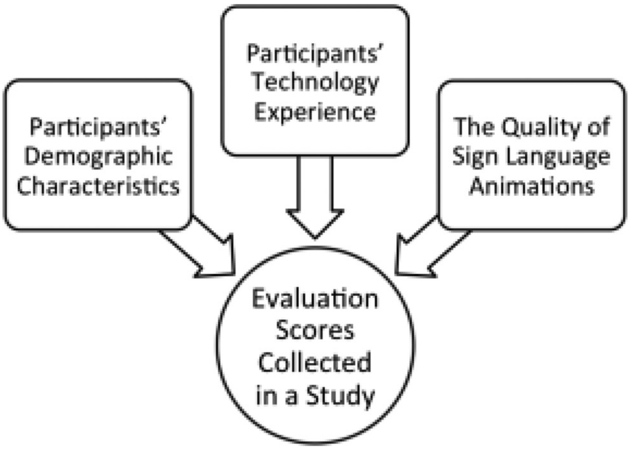
Participants' Demographic Characteristics
Participants' Technology Experience
The Quality of Sign Language Animations
Evaluation Scores Collected in a Study

Fig. 13. Graphical illustration of the factors that may affect the evaluation scores collected in a study in which deaf participants evaluate sign language animation.

—***Subjective and HomeASL.*** We observed that whether a participant used ASL at home was also a factor with a negative relationship to their subjective score for ASL animations. We can speculate that this might also be a case of "higher standards," that is, frequent ASL users may be harsher critics of ASL animation quality.

—***Subjective and SchoolType.*** While the SchoolType variable was important in both Comprehension Model 2 and in Subjective Model 2, the *direction of the relationship is reversed*. We observed that attending a residential school had a positive relationship with Comprehension scores, but it had a negative relationship with Subjective scores. We note that it is reasonable that an independent variable may have opposite relationship with each of our dependent variables: Prior research has found low correlation between a participant's subjective score for an animation and his/her comprehension score for it [Huenerfauth at al. 2008].

6.2. How Much Variance in the Dependent Variables is Explained by the Stimuli

When examining the models presented in Section 6, some readers may note that the R-squared values of the models were relatively modest (<0.4). This result was not surprising, given that those models are predicting users' comprehension and subjective scores based only on their demographic and experience/attitude characteristics. Section 2 described how prior ASL animation researchers generally assume that the value of such scores is based *upon the difference in quality of the animation stimuli that are shown to participants*. As illustrated in Figure 13, our preceding regression analysis suggests that participants' demographic characteristics and technology experience may also explain some of the resulting evaluation scores.

Perhaps counterintuitively, the models in Section 6 did **not** include a variable that indicated which type of stimulus was shown, for example, the specific message nor the specific avatar technology, which are two variables that presumably relate to a participant's evaluation scores. Instead, we intentionally examined whether we could construct regression models of the variance in evaluation scores based only on demographic characteristics and technology experience/attitude of the participants. The rationale for that decision was that our focus has been on the degree to which the demographic and technology-experience characteristics influence the final scores. However, now for purposes of comparison, it would be useful to consider how much of the overall

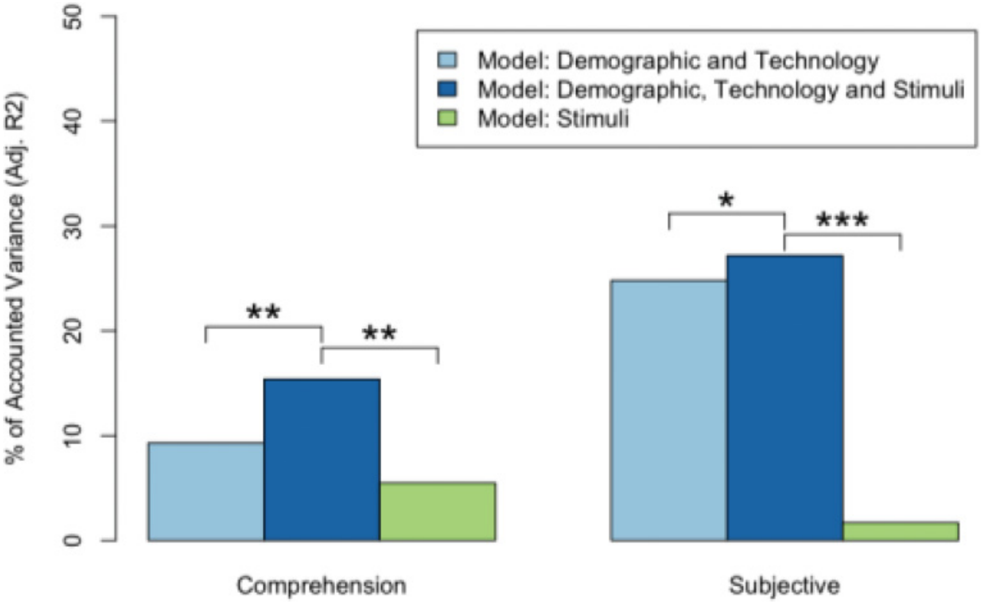

Fig. 14. Comparison of the Comprehension and Subjective Models from Section X to new versions of those models that include the AnimationType and Message variables. (Significance codes: 0 "***" 0.001 "**" 0.01 "*" 0.05.)

dependent score variance would be explained if we were to introduce two additional independent variables into our model, namely,

AnimationType: Which of the three types of avatar animation platform produced the animation that was shown? (EMBR, JASigning, VCOM)

StoryCode: Which of the three ASL passages was displayed to the participant in that stimulus? (N2, W2, and Y3) Note: These values are codenames for specific passages in ASL that serve as the script of animation that was shown.

By considering these two independent variables, we can examine how much of the R-squared variance of Comprehension scores and Subjective scores can be explained. Further, we can create models with different subsets of variables, and we can perform a statistical comparison (ANOVA) to see which are better.

In order to accommodate this new use of AnimationType and StoryCode as independent variables, it was necessary to make a modification to our data handling prior to modeling. Whereas the models shown in Section 6 were based on the average comprehension or subjective scores for each human participant (averaged across all three animations that had been displayed to that participant), in order to consider the AnimationType and StoryCode of the individual animations that were seen, we had to consider the comprehension or subjective scores for each animation individually. Instead of producing a single "comprehension" or "subjective" score for each participant in the study (as was done in Section 6), we now produce three comprehension scores and three subjective scores for each participant (one score for each of the three animations that the participant saw during their session). This change allows us to introduce the AnimationType and StoryCode variables into the regression modeling, since those stimuli-related variables are specific to an individual animation stimulus that was displayed. Since we are modeling a somewhat different dataset, the R-squared values for the "Model: Demographic and Technology" model displayed in Figure 14 differ from those shown for the "Model 2: Demographic and Technology" model displayed previously in Figure 11. Although both models contain an identical set of independent variables, the R-squared values displayed in the two figures differ, due this difference in how the data was handled prior to regression modeling.

Figure 14 presents the R-squared variance for three models of Comprehension scores and three models of Subjective scores:

—"Model: Demographic and Technology" is a multiple-regression model consisting of the set of variables selected in the "Model 2" listed in Tables III and IV in Section 6. (Note that the set of variables in the Comprehension model differs from the set used in the Subjective model.)

—"Model: Demographic, Technology, and Stimuli" contains this same set of variables as previously, with the addition of two variables: AnimationType and StoryCode.

—"Model: Stimuli" is a multiple-regression model consisting of only two independent variables: AnimationType and StoryCode.

Not surprisingly, adding additional variables, which describe the quality of the animation stimuli presented in the study, to the model allowed us to explain more of the variance: In Figure 14, note the lower R-squared value of the "Model: Demographic and Technology" bars, as compared to the "Model: Demographic, Technology, and Stimuli" bars, which include the two additional variables: AnimationType and StoryCode. Thus, as was suggested by the diagram presented in Figure 13, all three sets of independent variables (Demographic, Technology Experience/Attitude, and Stimuli Quality) have a relationship on the evaluation scores collected in a model.

The most relevant comparison is to consider the difference in R-squared value between the model trained on "Demographic, Technology, and Stimuli" variables and the one trained on "Stimuli" variables only. Here, we observe a significant higher R-squared value for the model with more variables. This is the key test of whether adding demographic and technology characteristics to the model can allow us to explain more variance in participant scores, as compared to a model based only on the characteristics of the stimuli that were presented.

A notable aspect of Figure 14 is the rather low R-squared value of the model trained on "Stimuli" variables only. This is somewhat counterintuitive: Most researchers might assume that what is primarily being measured by the comprehension or subjective questions in a study is the quality of the stimuli. Here, we can see that while such variables can explain part of the variance, in this case, a larger share of the variance was explained by the individual participants' demographic and technology-experience/attitude characteristics.

This finding suggests the importance of counterbalancing in the design of experimental studies evaluating animations of sign language: Specifically, if researchers are comparing alternative versions/platforms of animations, it may be prudent to use a study design in which each individual participant views and evaluates equal proportions of animations of each type. Of course, we must qualify this finding: In this study, while the three animations were produced by different animation platforms, all of them were of a somewhat similar level of quality (i.e., each including face/head movements and hand movements crafted by experts). We speculate that in a study with animation stimuli that varied more widely in their quality, we might have found that the "Stimuli" variables account for a greater share of the variance in Comprehension and Subjective scores.

7. STUDY #2: EXPERIMENTAL EVALUATION OF RELATIONSHIPS

While the regression model analysis presented in Section 6 has identified some suggestive relationships between demographic and technology-experience characteristics of users and their comprehension and subjective scores in a study, it was not an experimental study design. Thus, although that study may have suggested some hypotheses, the study did not formally evaluate any.

We therefore conducted a follow-up experimental study (referred to as "Study #2" in this article) with 57 additional participants evaluating animations of ASL. This study allowed us to formally evaluate the following five hypotheses, which are based on the results of our initial regression study, as summarized in Section 6.1:

$\mathbf{C}_{\mathbf{SchoolType}}$: When considering the comprehension-question response accuracy of participants evaluating ASL animations, those participants who attended residential or daytime schools for deaf students will have significantly *higher* scores than those participants who attended mainstream schools.

$\mathbf{S}_{\mathbf{HomeASL}}$: When considering the Subjective evaluation responses of participants evaluating ASL animations, those participants who use ASL at Home will have significantly *lower* scores.

$\mathbf{S}_{\mathbf{SchoolType}}$: . . . those participants who attended a residential or daytime school for deaf children will have significantly *lower* scores.

$\mathbf{S}_{\mathbf{MediaSharing}}$: . . . those participants with MediaSharing subscale scores indicating media sharing behaviors occurring more than once per month will have significantly *lower* scores.

$\mathbf{S}_{\mathbf{AnimationAttitude}}$: . . . those participants with AnimationAttitude subscale scores indicating an overall negative attitude will have significantly *lower* scores.

The first hypothesis relates to the comprehension-question response accuracy of participants in a study, and thus, the "C" in the codename of the hypothesis refers to "comprehension." The remaining four hypotheses relate to the subjective evaluation scores of participants; so, the "S" in the codename of these hypotheses refers to "subjective." Lower subjective scores indicate more negative subjective judgments.

For the $\mathrm{C}_{\mathrm{SchoolType}}$, $\mathrm{S}_{\mathrm{HomeASL}}$, and $\mathrm{S}_{\mathrm{SchoolType}}$ hypotheses, since the HomeASL and SchoolType variables have discrete values, it is straightforward to partition participants according to their responses to questions about these demographic characteristics. For the final two hypotheses, it was necessary to select threshold values in order to partition participants based on their score for the MediaSharing subscale or AnimationAttitude subscale on the technology-experience/attitude questionnaire. The rationales for selecting these threshold values are as follows:

- On the MediaSharing subscale, which consists of the average of responses to four questions about the individual's use of media and video online, if a participant responds "Never" to an individual question, this is registered as a value of 1 for that question. If the participant responds "Once per month" as the answer to a question, the response is registered as a 2, and if they select a response indicating greater frequency, for example, weekly or daily, the values are higher. Finally, the responses to these four questions are averaged together to produce the MediaSharing subscale value for that participant. We decided to partition those individuals with MediaSharing subscale values below 2.5 from those with higher scores, in order to differentiate between individuals who primarily selected "Never" or "Once per Month" responses and those individuals who selected responses indicating greater frequency of use of media or video online.
- On the AnimationAttitude subscale, which consists of the average of responses to six Likert items, a response of "Neither Agree nor Disagree" to any item is registered as a value of 3 for that item. Responses of "Disagree" or "Strongly Disagree" are 2

and 1, and responses of "Agree" or "Strongly Agree" are 4 or 5. Finally, the values for all of the individual items are averaged to produce the AnimationAttitude subscale score. Thus, we decided to partition those individuals who score below 3 on the subscale from those individuals with higher scores, in order to differentiate between individuals with negative or positive responses.

7.1. Participants and Stimuli in Study #2

In Study #2, Deaf researchers (all fluent ASL signers) recruited participants and conducted the data-collection sessions, with similar channels of online, in-person, and social networking advertisement used as in Study #1. A total of 57 participants were recruited to evaluate a set of ASL animations by responding to comprehension questions and subjective evaluation questions. The participants in this study responded to the demographic questionnaire and an abbreviated version of the technology-experience/attitude questionnaire (consisting of only the MediaSharing and AnimationAttitude question items).

A total of 57 people participated in the study, where 38 participants self-identified as deaf/Deaf and 19 as hard-of-hearing. Of our participants in the study, 13 had attended a residential school for deaf students, and 9, a daytime school for deaf students. 35 participants had learned ASL prior to age 5, and the remaining 16 had been using ASL for an average of 10 years. There were 32 men and 25 women of ages 18–32 (average age 22.3).

The animation stimuli shown in Study #2 were somewhat different than those shown in Study #1. In Study #2, all of the animations were produced using the EMBR animation system, with the facial expressions based on computer vision analysis of video-recordings of human ASL signers, as discussed in Kacorri [2016]. The script for the animations consisted of 10 of the ASL passages released to the research community in Huenerfauth and Kacorri [2014] to serve as a standardized testing stimuli set for conducting evaluations of ASL animations; specifically, passages with codenames N2, N5, R3, R9, T3, T4, W1, W2, Y3, AND Y4 were used in Study #2. In comparison, the stimuli in Study #1 were produced using three different animation platforms (EMBR, VCOM, and JASigning) and only three of these standard stimuli passages (N2, W2, and Y3).

As was done in Study #1, at the beginning of the study, participants viewed a sample animation, to familiarize them with the experiment and the questions they would be asked about each animation. (The Sample animation used a different stimulus than the other 10 animations shown during Study #2.) After viewing each of the 10 stimuli animations, participants answered subjective and comprehension questions, as they had done in Study #1.

7.2. Results of Study #2

To evaluate each of the five hypotheses in this study, we partitioned the participants in the study four different ways, according to each of the four variables:

HomeASL: Participants were partitioned into two groups: Those who answered "yes" to the question as to whether they used ASL at Home and those who did not.

SchoolType: Participants were partitioned into two groups: Those who attended a residential or daytime school for deaf children and those who attended a mainstream school.

MediaSharing: Participants were partitioned into two groups: Those with MediaSharing subscale scores below 2.5 and those with scores of 2.5 or above.

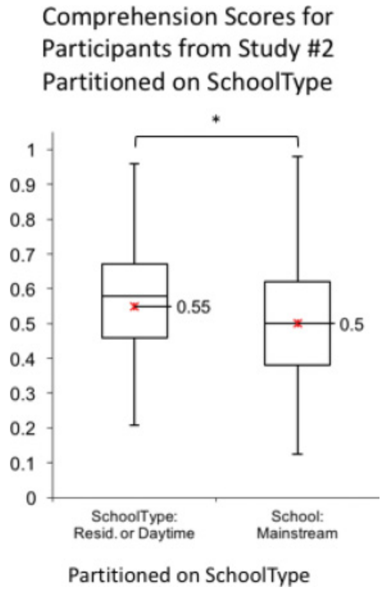

Fig. 15. Comprehension questions scores for those participants in Study #2 who attended a residential or daytime school for deaf students and those who attended a mainstream school. (Significance codes: 0.01 "*" 0.05.)

AnimationAttitude: Participants were partitioned into two groups: Those with AnimationAttitude subscale scores below 3 and those with scores of 3 or above.

Based on each of these partitions of the participants in Study #2, we performed comparisons of the average comprehension-question scores or average subjective scores for each animation, to evaluate each of the five hypotheses listed previously. Figure 15 presents the results for comprehension scores for the two partitions of the SchoolType variable. A t-test was used to compare the two groups of responses for the comprehension-question response accuracy; statistically significant differences are marked with an asterisk ($*$) in the figure. In the box plots in both Figures 15 and 16, the box represents the upper and lower quartile of scores, the midline represents the median value, the X indicates the mean (which is labeled with its value), and the whiskers represent the minimum and maximum values.

Figure 16 presents the results for subjective scores for all four methods of partitioning the participants in Study #2 (based on the four variables HomeASL, SchoolType, MediaSharing, and AnimationAttitude). A Mann-Whitney U test was used to compare the groups of Subjective responses: nonparametric tests are necessary for scalar-response data that are not normally distributed.

Based on the results shown in Figure 15, Hypothesis $C_{\text{SchoolType}}$ was supported; that is, those participants who attended a residential or daytime school for deaf students had higher comprehension-question response accuracy scores when viewing animations of ASL (as compared to participants who attended a mainstream school).

Based on the results shown in Figure 16, we can draw the following conclusions:

—Hypothesis S_{HomeASL} was not supported; that is, we did not observe a statistically significant difference between those participants who used ASL at home, as compared to participants who did not use ASL at home. Thus, this relationship that had been

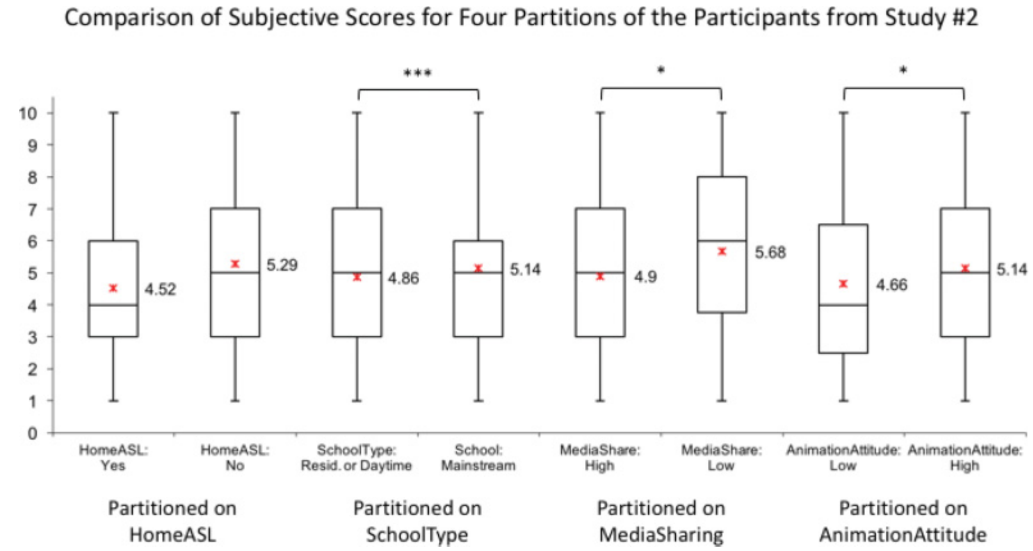

Fig. 16. Subjective scores for all four partitions of the participants in Study #2, along each variable: HomeASL, SchoolType, MediaSharing, or AnimationAttitude. (Significance codes: 0 "***" 0.001 "**" 0.01 "*" 0.05.)

suggested by the regression modeling results from Study #1 was not supported based on the experimental analysis during Study #2.

—Hypothesis $S_{SchoolType}$ was supported; that is, those participants who attended a daytime or residential school for deaf students had lower subjective response scores when viewing animations of ASL, as compared to those participants who had attended a mainstream school. We speculate that they were more critical judges of the ASL quality, due to their increased use of ASL.

—Hypothesis $S_{MediaSharing}$ was supported; that is, those participants with high MediaSharing subscale scores had lower subjective response scores when viewing animations of ASL. We speculate that the more frequent users of technology were more critical when evaluating animations.

—Hypothesis $S_{AnimationAttitude}$ was supported; that is, those participants with low AnimationAttitude subscale scores had lower subjective response scores when viewing animations of ASL. We speculate that those individuals with negative attitudes about animation technology in general had more negative scores when evaluating specific stimuli.

Overall, this experimental study has confirmed most of the relationships that were informally suggested by the earlier regression analysis based on Study #1 in Section 6. In the case of Comprehension scores, we observed a significant difference in scores for participants, depending on their response to the SchoolType question. In the case of Subjective scores, we observed significant differences for the SchoolType, MediaSharing, and AnimationAttitude variables, but we did not observe a significant difference when we partitioned the participants using the HomeASL variable.

8. TEXT-BASED RESPONSES AND FEEDBACK FROM PARTICIPANTS

Before presenting the major conclusions of this article in Section 9, we wanted to use this space to briefly summarize some data that had not been previously analyzed from our original Study #1. Specifically, our questionnaire given to those 62 participants included three questions where participants could give text responses:

—Could you suggest ways that computer animations of sign language could be used?
—Could you list what you liked about the computer animations that you saw today?
—Could you list some of the things that should be improved?

Because these participant responses were not included in the regression analysis in Section 6, we delayed discussing them until this section so as not to interrupt the narrative progression of earlier sections of this article. Although these responses were not part of our analysis of the relationship between participant characteristics and comprehension or subjective scores, this feedback from users suggests potential applications of this technology and recommendations about how to improve it.

Our methodology for analyzing this text data from the 62 participants was as follows: Upon examining the responses to the questions, we noted that participants tended to respond to questions with text that addressed multiple questions simultaneously. For instance, some respondents mentioned aspects that could be improved as part of their response to the question about what they "liked" about the animations. For this reason, we decided to treat all of the text responses from each participant as a single text. To look for patterns in the response text, we used an open coding strategy. After an initial pass of coding, a second round was conducted to look for groupings of the codes and to improve coding consistency. Finally, a second researcher reexamined the text and the coding, and a consensus was reached prior to summarizing the text comments in the following.

Many of the responses about possible applications of sign language animation technology fell into a few frequent categories: Of the 62 participants, nine mentioned use of sign language animations in public transportation (e.g., airports, train stations), nine mentioned use in public spaces or for public announcements (e.g., shopping malls), nine mentioned educational applications (e.g., ASL dictionaries or software to demonstrate signs to children), seven mentioned use in entertainment programs (e.g., as a form of captioning for movies or television), five mentioned use on websites (e.g., as a language option that users could select), and five mentioned use in restaurants (e.g., when ordering inside a fast food restaurant or at a drive-through window). Other participants recommended more specific environments in which such technology could support communication (the number of participants who mentioned each is shown in parentheses), including police stations (two), "911" emergency calls (one), doctors offices or hospitals (one), grocery stores (one), or welcome centers for institutions (one). While many of the preceding suggestions might use a public display screen showing an animation or someone viewing animation on a personal computer, three participants specifically mentioned viewing such animations on a mobile telephone, including for conveying a voicemail message or when using GPS directions. Two participants mentioned that they would like to see this technology appear in video games.

Our participants had a wide variety of opinions about the quality of sign language animation technology and whether it should be deployed in future accessibility applications. Several had positive reactions to the technology:

—12 participants commented generally about the future potential of the technology, for example, "potential usage in the future and cool idea," "good starting point with the avatars," "they seemed like a great beginning," "it was fascinating to see ASL being signed; it does seem understandable," "it's interesting to see animation signing," and "it amazed me that they can do sign language."
—Nine participants commented how the animations were understandable, often expressing surprise about this, for example, "comfortable understanding," "better [than] I expected—realistic and somewhat understandable," "some animations were surprisingly understandable," "it does seem understandable," and "some of them were very clear [in] their signing."

—The variety of application areas that participants recommended for this technology (listed previously) is another indication of positive subjective views.

Negative comments from participants often consisted of a general rejection of the technology or a mention of specific situations in which it would not be suitable:

—Four participants said that they would not want to see sign language animations used in any context, for example, commenting: "worthless to be used because they were difficult to understand," "no, I won't suggest it," "I dislike it because they are too robotic and they are useless," and (in response to a question of how the technology could be used, replying) "not at all."

—An additional four participants mentioned that they would not want to see it used in specific applications: For instance, one said "not for relay" (indicating that it should not be used for telephone relay services), and another mentioned that they did not believe it was appropriate for "1:1 contact" between two people as a communication aid.

—Some were concerned about replacing human interpreters with animations; three mentioned a preference for human signing, for example, commenting: "It seems useful, but human are better," "I'd prefer human sign language instead of animation," and "use live person to interpret and give them a job."

Our laboratory has never advocated for use of animation technology as a replacement for interpreters; given the state of the art of automatic machine translation technology for sign languages, we are concerned that such a suggestion might lead to reduced accessibility for people who are deaf. Instead, Section 1 describes our focus on providing ASL on websites or other information sources for which information is in a (less accessible) written form and where a human interpreter is not available. In Huenerfauth and Hanson [2009], we discuss the ethical responsibilities of researchers working on sign language animation technologies to communicate the capabilities of this technology clearly to avoid its premature usage to avoid reducing the quality of accessibility currently provided through other means, such as human interpreters. Given this context, we were surprised that seven participants mentioned using sign language animation technologies as an alternative to interpreters—although they generally qualified this suggestion by mentioning that it might be suitable when a human interpreter is not available, when the need for interpretation is unexpected, or in a context that is not amenable for a human interpreter. Participants commented: "areas that are not terp friendly, or maybe while waiting for a terp," "to be used when there's no interpreter available," "impromptu interpreter in situations like . . . an app with voice recognition."

In regard to specific aspects of the animations that they liked or that needed improvement, participants expressed conflicting opinions on the smoothness of motion, facial expression quality, signing speed, and appearance of the characters.

—*Smoothness of movement:* While one participant had a positive comment about the smoothness of the animated character's movements ("I was impressed with the ability to make them smooth"), a majority (32) commented that the animations should be smoother, for example, asking for "smooth signing," "less robotic," "fluid with their motion," and "less choppiness."

—*Facial expression:* 18 participants indicated that the facial expression, lip movements, or eye movements of the characters needed to be improved, for example, commenting "improve facial expression and add emotion," "mouth movements added," or "facial expression involving eyes and mouth." Five participants had neutral-to-positive comments about facial expression (although generally only weakly positive, e.g., "okay").

—*Speed:* Participants disagreed about the speed of the animations. Two commented that the speed was appropriate, and five commented that the animations should be slower or faster. One participant specifically mentioned that it would be nice for the speed to be adjustable by the person viewing it.

—*Appearance:* Participants also disagreed about the appearance of the characters, specifically the background color, the clothing of the characters, the skin color of the characters, and the apparent gender of the characters. Eight participants had positive comments, for example, "they dressed good," "(good) fashion," "I liked the background color," "solid background, solid clothes, no distractions," "I liked how they used different skin colors and genders." Other participants recommended changes in the characters' appearance, for example, "the color of the background can be green like how you see an interpreter thru videophone," "change different backgrounds," "clothes should be bright colors; less jacket," and "maybe dark clothes for light skin, light clothes for dark skin."

Overall, based on the feedback comments from participants, the smoothness of the animation movement and (to a somewhat lesser degree) the quality of the facial expressions should be considered high-priority concerns for ASL animation researchers. Given the differences in opinions about animation speed and character appearance, researchers may want to consider making these aspects of computer animations adjustable or customizable by end users, to suit their preferences.

9. CONCLUSIONS AND FUTURE WORK

As described in Section 1, the long-term goal of our research is to investigate the design of software to automatically synthesize animations of sign language from a simple script of the desired message. This automatic animation-creation technology would make it easier to maintain and update information online in the form of sign language. As part of this research agenda, we are interested in understanding how to best conduct studies to evaluate the quality of such software; such methodological research is needed to ensure progress in the field. The findings of the studies presented in this article will affect the set of demographic and technology-experience/attitude questions we ask participants in future work. Thus, one contribution of this research is a deeper understanding of the relationship between participant characteristics and evaluation scores in this field. Specifically, we found that the following variables were most important in explaining variance in comprehension and subjective scores of sign language animations:

—**SchoolType:** Assessed with a single multiple-choice question.
—**HomeASL:** Assessed with a single polar (yes-or-no) question.
—**MediaSharing:** Assessed with four scalar-response items indicating frequency of different activities, from Rosen et al. [2013].
—**AnimationAttitude:** Assessed with six Likert agreement items.

While we have noted other variables that were present in some of the regression models presented in Section 6, the preceding four items correspond to the most important factors (as discussed in Section 6.1). Collecting this abbreviated set of variables may be useful for researchers interested in minimizing the amount of study time spent collecting demographic and technology-experience/attitude data. Of course, we anticipate researchers may continue collecting and reporting other demographic data about their participants (e.g., age or gender), but our survey of prior work in Section 2.1 suggests that few current sign language animation researchers regularly collect and report these preceding four items.

While all four of these variables were identified as having relationships to Comprehension and Subjective scores during the regression modeling in Section 6, during our subsequent experiment study in Section 7, we were unable to confirm the relationship between HomeASL and Subjective Scores. Despite this nonsignificant result in Section 7, we still recommend that future researchers ask participants about the HomeASL question and report the responses of their participants in publications. Our rationale for continuing to recommend this variable is twofold:

—The tradition of children attending residential or daytime schools specifically for deaf children is somewhat specific to the educational system in the U.S., and researchers from other countries who are evaluating sign language animation technologies may not find the variable of SchoolType as relevant for their population. Furthermore, there has been a trend over the past two decades in the U.S. for more deaf students to attend mainstream educational programs (instead of schools specifically for deaf children), thus, the value of the variable of SchoolType to distinguish participants with higher ASL skill may change over time, due to these changing educational trends.

—The HomeASL question is not very time-consuming to collect from participants since it is a brief yes-or-no question. Given the potential considerations about the SchoolType variable mentioned previously, we speculate that the HomeASL variable may be a possible replacement variable that may indicate individuals with greater ASL usage.

In prior work, we have released stimuli and evaluation questions to the research community, in order to promote replicability and comparison of results across studies [Huenerfauth and Kacorri 2014]. We have made use of these sets of stimuli in the studies presented in this article. In a similar manner, we hope to further contribute to research replicability and consistency of evaluation in our research community by sharing the survey questions (both English text and the ASL videos) used in the studies reported in this article, which can be found in the online appendix to this article in the ACM Digital Library.

Through collection and publishing of these demographic and technology-experience/attitude characteristics of participants by researchers evaluating sign language animation technologies, we anticipate that it may be easier to compare research results across publications. We also believe that these factors may be useful for researchers to consider if they are balancing or matching participants across treatment conditions in a study.

Compared to prior non-online studies evaluating sign language animation, the studies presented in this article were relatively large (N = 62 for Study #1, N = 57 for Study #2). However, in future work, it would be useful to recruit more participants from the Deaf community in another geographic area (outside Rochester, NY), to ensure that the relationships observed in the current study are preserved.

Furthermore, in future work, we are interested in exploring the variable of Age. This variable was not selected by the exhaustive all-subsets model comparison in Section 6, but only 10% of our 62 participants in that study were over age 43. In future work, we would like to conduct additional targeted recruitment of older participants. As we have learned when conducting the two studies described in this article, it was relatively more time-consuming to recruit older participants; so, this must be factored into the data-collection timeline in future work.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

We are grateful to Sandra Sidler-Miserez, who helped produce ASL animations, and to Evans Seraphin, who assisted in the recruitment of participants and in data collection.

REFERENCES

Bonnie B. Blanchfield, Jacob J. Feldman, Jennifer L. Dunbar, and Eric N. Gardner. 2001. The severely to profoundly hearing-impaired population in the United States: Prevalence estimates and demographics. *J. Am. Acad. Audiol.* 12 (2001), 183–189.

Center for Research and Education on Aging and Technology Enhancement (CREATE). 2015. Resources. Retrieved May 6, 2015 from http://create-center.gatech.edu/resources.php.

Michael Crabb and Vicki L. Hanson. 2014. Age, technology usage, and cognitive characteristics in relation to perceived disorientation and reported website ease of use. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'14)*. ACM, New York, 193–200. DOI:10.1145/2661334.2661356

Sarah Ebling and John Glauert. 2015. Building a Swiss German sign language avatar with JASigning and evaluating it among the deaf community. *Univ. Access Inf. Soc.* 1–11.

P. E. Mohr, J. J. Feldman, J. L. Dunbar, A. McConkey-Robbins, J. K. Niparko, R. K. Rittenhouse, and M. W. Skinner. 2000. The societal costs of severe to profound hearing loss in the United States. *Int. J. Technol. Assess. Health Care* 16, 4 (2000), 1120–1135.

John Fox and Georges Monette. 1992. Generalized collinearity diagnostics. *JASA 87* (1992), 178–183.

Andrew Gelman. 2008. Scaling regression inputs by dividing by two standard deviations. *Stat Med* 27, 15 (1992), 2865–2873.

Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The *SignCom* system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 6 (October 2011). DOI:10.1145/2030365.2030371

Ulrike Grömping. 2006. Relative importance for linear regression in R: The package relaimpo. *J. Stat. Softw.* 17, 1 (2006), 1–27.

Thomas Hanke. 2001. *ViSiCAST Deliverable D5–1: Interface Definitions.* Technical Report. ViSiCAST project. Retrieved February 15, 2016 from http://www.visicast.co.uk/members/milestones/D5-1rev1.pdf.

Kyle Hayward, Nicoletta Adamo-Villani, and Jason Lestina. 2010. A computer animation system for creating deaf-accessible math and science curriculum materials. In *Proceedings of Eurographics'10*.

Alexis Heloir, Quan Nguyen, and Michael Kipp. 2011. Signing avatars: A feasibility study. In *Proceedings of the 2nd International Workshop on Sign Language Translation and Avatar Technology*.

Judith A. Holt, Sue Hotto, and Kevin Cole. 1994. Demographic aspects of hearing impairment: Questions and answers (3rd ed.). Center for Assessment and Demographic Studies, Gallaudet University. Retrieved April 3, 2016 from https://research.gallaudet.edu/Demographics/factsheet.php.

M. Huenerfauth and V. Hanson. 2009. Sign language in the interface: Access for deaf signers. In *Universal Access Handbook*, C. Stephanidis (Ed.). Lawrence Erlbaum Associates, Mahwah, NJ. 38.1–38.18.

Matt Huenerfauth and Hernisa Kacorri. 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'14)*.

Matt Huenerfauth and Hernisa Kacorri. 2015. Best practices for conducting evaluations of sign language animation. In *Proceedings of the 30th Annual International Technology and Persons with Disabilities Conference (CSUN'15)*. Scientific/Research Track.

Matt Huenerfauth, Liming Zhao, Erdan Gu, and Jan Allbeck. 2008. Evaluation of American Sign Language generation by native ASL signers. *ACM Trans. Access. Comput.* 1, 1, Article 3 (May 2008), 27 pages. DOI:10.1145/1361203.1361206

International Standards Organization (ISO). 2004. ISO/IEC 14496–2:2004: Information technology—Coding of audio-visual objects – Part 2: Visual.

Vince Jennings, Ralph Elliott, Richard Kennaway, and John Glauert. 2010. Requirements for a signing avatar. In *Proceedings of the Workshop on Corpora and Sign Language Technologies (CSLT), LREC*. 33–136.

Hernisa Kacorri. 2016. *Data-Driven Synthesis and Evaluation of Syntactic Facial Expressions in American Sign Language*. Dissertation. Computer Science, The Graduate Center, City University of New York.

Hernisa Kacorri and Matt Huenerfauth. 2014. Implementation and evaluation of animation controls sufficient for conveying ASL facial expressions. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'14)*. ACM, New York, 261–262. DOI:10.1145/2661334.2661387

Hernisa Kacorri and Matt Huenerfauth. 2015. Comparison of finite-repertoire and data-driven facial expressions for sign language avatars. In *Universal Access in Human-Computer Interaction. Access to Interaction.* Springer International Publishing, 393–403.

Hernisa Kacorri, Matt Huenerfauth, Sarah Ebling, Kasmira Patel, and Mackenzie Willard. 2015. Demographic and experiential factors influencing acceptance of sign language animation by deaf users. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'15)*. ACM, New York, 147–154. DOI:http://dx.doi.org/10.1145/2700648.2809860

Michael A. Karchmer and Ross E. Mitchell. 2004. Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies*, 4, 2 (2004), 138–163.

Richard Kennaway, John R. Glauert, and Inge Zwitserlood. 2007. Providing signed content on the Internet by synthesized animation. *ACM Trans. Comput.-Human Interact.* 14, 3 (Sept 2007), 15. DOI:10.1145/1279700.1279705

Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11). ACM*, New York, 107–114. DOI:10.1145/2049536.2049557

Richard H. Lindeman, Peter F. Merenda, and Ruth Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis*. No. 519.535 L743, Scott Foresman, Glenview, IL.

Thomas Lumley and A. Miller. 2009. *Leaps: Regression Subset Selection. R package version 2.9.*

Ross M. Mitchell, Travas A. Young, Bellamie Bachleda, and Michael Karchmer. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang. Studies* 6, 3 (2006), 306–335.

Kgatlhego A. Moemedi. 2010. *Rendering an Avatar from Sign Writing Notation for Sign Language Animation*. Doctoral dissertation. University of the Western Cape.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Jan Henning. 1989. *HamNoSys: v2.0: Hamburg Notation System for Sign Languages: An Introductory Guide*. Signum, Hamburg, Germany.

Larry D. Rosen, Kelly Whaling, L. Mark Carrier, Nancy A. Cheever, and J. Rokkum. 2013. The media and technology usage and attributes scale: An empirical investigation. *Comput. Human Behav.* 29, 6 (2013), 2501–2511.

Kimberly K. Speerschneider and Jason M. Bryer. 2013. likert: An r package for visualizing and analyzing likert-based items. In *Proceedings of the useR! Conference*.

Jessica J. Tran, Tressa W. Johnson, Joy Kim, Rafael Rodriguez, Sheri Yin, Eve A. Riskin, Richard E. Ladner, and Jacob O. Wobbrock. 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 (ASSETS'10)*. ACM, New York, 115–122. DOI:10.1145/1878803.1878825

Carol B. Traxler. 2000. The Stanford achievement test, 9th edition: National norming and performance standards for deaf and hard-of-hearing students. *J. Deaf Stud. Deaf Educ.* 5, 4 (2000), 337–348.

VCom3D. 2015. Homepage. Retrieved from http://www.vcom3d.com.

Margriet Verlinden, Corrie Tijsseling, and Han Frowein. 2001. Sign language on the WWW. In *Proceedings of 18th International Symposium on Human Factors in Telecommunication*.

Ou Yang, Kenichi Morimoto, and Noriaki Kuwahara. 2014. Evaluation of Chinese sign language animation for mammography inspection of hearing-impaired people. In *Proceedings of Advanced Applied Informatics (IIAI'14)*. IEEE, 831–836.

Received April 2016; accepted December 2016