Authors' Unofficial Copy of:
Kacorri, H., Huenerfauth, M., Ebling, S., Patel, K., & Willard, M. (2015, October). Demographic and Experiential Factors Influencing Acceptance of Sign Language Animation by Deaf Users. In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (pp. 147-154). ACM.

Demographic and Experiential Factors Influencing Acceptance of Sign Language Animation by Deaf Users

Hernisa Kacorri[1], Matt Huenerfauth[2], Sarah Ebling[3], Kasmira Patel[2], Mackenzie Willard[2]

[1]The Graduate Center, CUNY
365 Fifth Avenue
New York, NY 10016 USA
+1-212-817-8190
hkacorri@gc.cuny.edu

[2]Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14623 USA
+1-585-475-2700
{matt.huenerfauth, kxp3157, mtw2935}@rit.edu

[3]University of Zurich
Binzmühlestrasse 14
8050 Zurich Switzerland
+41-44-635-71-45
ebling@cl.uzh.ch

ABSTRACT
Technology to automatically synthesize linguistically accurate and natural-looking animations of American Sign Language (ASL) from an easy-to-update script would make it easier to add ASL content to websites and media, thereby increasing information accessibility for many people who are deaf. Researchers evaluate their sign language animation 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/attitude characteristics of participants and (b) the subjective and objective scores collected from them during the evaluation of sign language animation systems. This finding suggests that it would be important for researchers to collect and report these characteristics of their participants in publications about their studies, but there is currently no consensus in the field. We present a set of questions in ASL and English that can be used by researchers to measure these participant characteristics; reporting such data would enable researchers to better interpret and compare results from studies with different participant pools.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**] User Interfaces – *evaluation/methodology*; K.4.2 [**Computers and Society**]: Social Issues – *assistive technologies for persons with disabilities*.

General Terms
Design, Experimentation, Human Factors, Measurement.

Keywords
Accessibility Technology for People who are Deaf; American Sign Language; Animation; User Study.

1. INTRODUCTION
Access to understandable information on websites and other media is essential in many education, commerce, and social contexts, yet most online content is in the form of written language text. Many people who are deaf and hard-of-hearing experience reduced exposure to language during childhood or educational circumstances that may lead to lower written language literacy: The median literacy rate of deaf high school graduates in the U.S. is at the 4th-grade level [27]. (U.S. 4th-grade students are typically age 10.) There are over 500,000 people in the U.S. who use American Sign Language (ASL) as a primary language [22]. Given the linguistic differences between English and American Sign Language (ASL), it is possible to be fluent in one language but not the other. Thus, providing information online in ASL can make content more accessible to users with lower English literacy; furthermore, many Deaf[1] users may simply prefer to receive information content in their primary language of ASL.

While posting videos of human signers on websites may seem like a simple solution, such videos are difficult to update when information changes, and they do not enable just-in-time generation of website content from a user request. For this reason, several research groups, e.g., [9, 15, 18, 19, 29], are studying the development of software to automatically synthesize understandable animations of a virtual human performing sign language, based on an easy-to-update script as input. The key challenge is for this software to select the details of such animations so that they are linguistically accurate, understandable, and acceptable to users. Researchers generally evaluate the quality of their software by: generating animations using some current version of their software, setting up an experiment in which deaf participants view and evaluate the animations, and comparing the scores of animations produced using the software (to some baselines or to prior versions). However, the field lacks consensus about the set of demographic data that should be reported about the participants. Thus, it is difficult to compare the results across studies because some variation in comprehension or subjective evaluation scores in studies may be explained by demographic characteristics of the participants, rather than by true differences in the quality of the animations being evaluated.

The goal of this paper is to examine the use of demographic and technology-experience variables as predictors of participants' responses to (a) subjective measures of animation quality and (b) objective measures of comprehension of the content. We present 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-ii) relate to participants' responses (a-b).

[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.

ASSETS '15, October 26-28, 2015, Lisbon, Portugal

DOI: http://dx.doi.org/10.1145/2700648.2809860

The remainder of this paper is organized as follows: Section 2 presents related work on demographics and evaluation studies. Sections 3, 4, and 5 describe the methodology of our study, and sections 6 and 7 present the results and discussion. Finally, Section 8 summarizes our conclusions and future directions.

2. RELATED WORK

In this related work section, we examine how prior sign language animation researchers have considered demographic or attitude variables when conducting their studies. We 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 some researchers have conducted studies of how various demographic or health factors affect technology use and acceptance, e.g., [1, 2, 25], this section focuses on studies with deaf participants evaluating sign language animations.

2.1 Demographics in Prior Studies

The primary focus of this paper is how demographic characteristics or technology experience of participants in a study may affect the results collected. We therefore surveyed prior sign language animation studies to identify the types of participant characteristics or technology experience/attitudes that researchers reported. Our goal was to understand the diversity of participants in prior studies and the types of data that researchers commonly collect. While there are a few examples of published results where only the number of participants and how they are self-identified is reported, e.g., [23, 30], in general the trend in the field is to include more information about the sampled population. Table 1 presents examples of representative papers, similar patterns may be found when examining larger surveys of prior evaluation studies, e.g., [3, 11, 13, 16, 19].

Common characteristics reported in studies include the age range of participants, the gender ratio of participants, and the ratio of participants identifying as deaf/Deaf or hard-of-hearing (this is indicated by the term "describe" in Table 1). Studies vary in how they measure and report the level of sign-language skill of their participants, e.g., using "signing frequency" or "self-reported sign language skills." We also see variability in whether researchers ask about participants' exposure to technology; for example, [9] included questions about "computer expertise." Researchers in [29] noted that only those participants who were unfamiliar using the Internet had negative attitudes towards their avatar; the authors 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.*" 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"). When included, such questions provide insight into how participants may see this technology being applied, e.g. as an educational tool [9] or for giving information in public spaces [18].

Table 2 shows some values of the most commonly reported demographic characteristics from the papers in Table 1. We can see wide variation in the demographic characteristics of users in prior sign language animation 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, e.g., some described what language their participants preferred [18, 29] and others described how often they used signing [18].

Table 1: Demographic and Technology Experience Characteristics Reported in Example User Studies

Paper	*Demographic*	*Technology*	*Attitudes*
[9]	age, gender, describe, profession	computer expertise	animation usage
[18]	age, gender, describe, signing frequency, preferred language		animation usage
[29]	age, gender, preferred language		attitude to avatar
[6]	age, gender, describe, self-reported SL skills, location		attitude to avatar

Table 2: Demographic Profile of Participants in Prior Studies

Paper	*Age Range*	*Female: Male*	*Describe*	*Assessing Signing Skills*
[9]	35-50	4 : 1	Deaf	Deaf educators
[18]	16-66	"slightly less female"	"most were deaf, some were hard-of-hearing"	"all were good signers… all using signing on a daily basis"
[29]	20-53	5 : 4	deaf	Some had preference for sign language; others had no preference between signing or text.
[6]	19-56	18 : 7	17 deaf, 8 hearing	8 "good," 6 "very good," 11 "native/expert"

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 comprehension scores or subjective judgments of the participants?* Knowing the answer to this question 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 pre-disposed to have positive or negative evaluations of ASL animations). Thus, the goal of our study (sections 3 to 5) 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 (and thereby reported by researchers in their publications) to facilitate comparison of results across papers.

Some 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 [29]. However, due to the relatively small sample size of most prior studies, researchers rarely present quantitative results for sub-populations. 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

Kipp et al. [19] carried out the most comprehensive study to date with participants evaluating multiple sign language avatars; thus, in this section, we position our new study in relation to this most-closely related prior work. In focus groups in [19], eight native signers of German Sign Language were presented with six avatars signing content in different sign languages, and they commented on their quality. While researchers in [19] showed participants animations in American Sign Language and other unfamiliar languages, in our study described in section 4, participants were shown animations in a language in which they were fluent (ASL). Further, researchers in [19] showed participants some hand-animated avatars (produced through a painstaking process of carefully posing the character). 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. [19] 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, we include objective comprehension questions to measure participants' understanding. Self-reports of understanding typically have low correlation to a participant's accuracy at answering comprehension questions [13].

Notably, in both the focus group and the online survey, the authors observed higher scores in response 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. In our new study, we include a question about whether participants had previously seen computer animations of sign language (details section 3.2).

Participants in [19] 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.

While [19] collected some demographics (gender, age, deaf/hard-of-hearing/hearing, and profession), they did not analyze the data to look for relationships between these factors and the survey responses. Our new study includes a regression analysis to identify demographic and/or experience factors that related to the participants' subjective responses and comprehension scores.

Given the online modality of [19], there is a possibility that participants could have been more comfortable using the Internet than the general population. In our new study (sections 3 and 4), 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 participation of less technology-savvy individuals and to allow for us to confirm that participants met our study criteria (and were accurately reporting their demographic data).

3. COLLECT INDEPENDENT VARIABLES

The goal of our work is to examine 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 section explains the design of our questionnaire for recording these independent variables, which will be 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., [25].

While some researchers have explored the design of fully online surveys of deaf users containing both ASL and English, e.g., [26], our survey was conducted 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 pre-existing English surveys (section 3.2); 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 on our lab website: http://latlab.ist.rit.edu/assets2015

3.1 Demographic Questions

We selected demographic questions by assembling items that were asked in prior experimental studies, e.g. [12], and questions asked in studies surveyed in section 2. Below, we list the demographic questions, 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?

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? (residential school for deaf students, 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? (did not graduate high school, graduated high school, graduated college, have bachelor's degree, have 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)

Note: After collecting data from participants (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**.

3.2 Technology Experience and Attitudes

To measure participants' frequency of technology use, we adopted the InternetSearch and MediaSharing subscales from the *Media and Technology Usage and Attitudes Scale* [25]; 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 below) on computers, laptops, tablets, or mobile phones:

InternetSearch: Search the Internet for news. …for information. …for videos. …for images or photos.

MediaSharing: Watch TV shows, movies, etc. Watch video clips. Download media files from other people. Share your own.

Using the same scoring, we created an ASLChat subscale:

ASLChat: Have a signing (ASL) conversation with someone using a video phone. 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 [25], in which the score is the average of responses to individual statements listed below (Strongly agree = 5, Agree = 4, Neither agree no 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 [1], using identical Likert scoring as above.

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 [6, 18, 19]. 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 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. COLLECT DEPENDENT VARIABLES

Section 2.2 described how [19] 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, there is one type of “diversity” from [19] that we did preserve in our study design: We wanted the results of this study to be applicable 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 [15, 17, 28]. In addition, in the [19] study, 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 [11]. Specifically, we selected three stimuli (codenames N2, W2, and Y3) that had been rated as being the most understandable in an earlier study by [17]. Example stimuli from the current study may be viewed on our lab website here: http://latlab.ist.rit.edu/assets2015

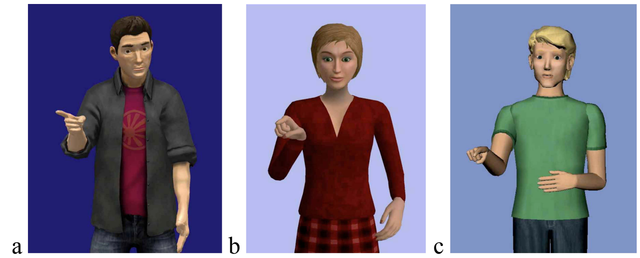

Figure 1: Screenshots from the three avatars shown in the study: (a) EMBR, (b) JASigning, (c) VCom3D.

EMBR: Animations shown in Figure 1(a) were generated using the open source EMBR platform [10] extended with ASL handshapes and detailed upper-face controls using the MPEG-4 Facial Animation standard [14]. Native ASL signers, who selected key-poses to define each sign in the lexicon, created the avatar’s hand movements. Video recordings of a native signer performing the stimulus automatically drove face and head movements. To extract the facial features and head pose from video of a human signer, we used Visage Face Tracker, an automatic face tracking software that provides MPEG-4 compatible output [16], which we converted into the script language supported by the EMBR platform, as described in [14].

JASigning: Animations shown in Figure 1(b) were produced using the free Java Avatar Signing (JASigning) system [15]. All signs in the stimuli were notated in the Hamburg Notation System (HamNoSys) [24] by a deaf researcher while consulting the video recordings of an ASL native signer performing the stimulus. HamNoSys, which serves as an input for JASigning, has around 200 symbols describing the components: handshape, hand position, location, and movement. Information about the non-manual components (e.g., eyebrow movement, eye gaze, and head movement) is included in the SiGML code [8], an XML representation for HamNoSys, but time-alignment of non-manuals with the manual signs requires careful adjustment, e.g., [3].

VCOM: Animations shown in Figure 1(c) were generated using the commercially available ASL authoring tool, VCom3D Sign Smith Studio [28], which allows users to produce animated ASL sentences by arranging a timeline of animated signs from a prebuilt or user-defined vocabulary. The software 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. The VCOM and EMBR animations shared similar hand movements.

During our study, after participants answered the demographic and technology-experience questions described in section 3, 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 three animations shown in the study.) Next, after viewing each of the three main animations, an onscreen 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 [11], a single “Comprehension” score for each animation can be calculated by averaging the scores of the four questions. Prior research, e.g., [12], has 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, the appropriate method for presenting comprehension questions and instructions, and other factors that we have considered in the design of this current study.

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 [11]. Questions (d-f) were inspired by [19]. To calculate a single "Subjective" score for each animation, the scalar response scores for the six questions were averaged.

5. RECRUITING & DATA COLLECTION

Prior research, e.g. [13], has discussed the advantages of having deaf researchers conduct experimental studies in ASL. In this study, a deaf researcher (co-author) 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 four-week data collection period, a short timeframe made possible due to the many people who are deaf and hard-of-hearing associated with RIT or in Rochester, NY. It was easier for us 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. 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 3 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 (average age 25.73). For participants over age 43 (average age 53.14), there were 4 men and 2 women who learned ASL prior to age 9, 5 self-reported to be deaf/Deaf and 1 hard-of-hearing.

6. ANALYSIS AND RESULTS

The goal of our analysis was to examine how demographic factors relate to participants' responses to subjective and comprehension questions about ASL animations. In addition, we wanted to know whether variance in scores could be explained by participants' technology experience and attitudes. We therefore used multiple regression to analyze the data. 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. Many researchers, e.g., [2], follow the recommendation of [5] that continuous-value variables be normalized by dividing the individual participant metrics by two times the group standard deviation, to facilitate easier comparison among coefficients of scalar and binary predictors. We have followed this procedure for all of the continuous independent variables in this study.

We 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 paper 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. We used the 'leaps' package [21] to build models of all possible subsets of features to identify the model with the highest adjusted R-squared value (indicating 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 [4].

Table 3 summarizes the regression analysis for Comprehension.[2] In Model 1 (demographic variables only), the type of school that the participant attended had the largest coefficient (see "Estimate" column): attending a residential school for deaf students had a positive relationship with the participant's success at answering comprehension questions. Model 2 contained both demographic and technology 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 demographic and technology characteristics of each participant, e.g., their use of the Internet, positive attitude towards technology, and video game exposure. (Section 7 includes additional discussion of these factors.)

Table 4 summarizes the regression analysis for Subjective scores. In Model 1 (demographic variables only), 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.

[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 3: 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

Table 4: 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 **

Figure 2 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%.

It is not surprising that the adjusted R^2 values of the models are relatively low (<0.4), given that we predict users' comprehension and subjective scores based only on their demographic and experience/attitude characteristics. Section 2 described how prior

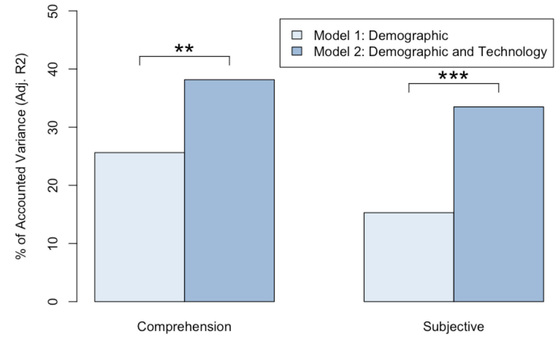

Figure 2: Regression model comparison summary. (Significance codes: 0 '*' 0.001 '**' 0.01)**

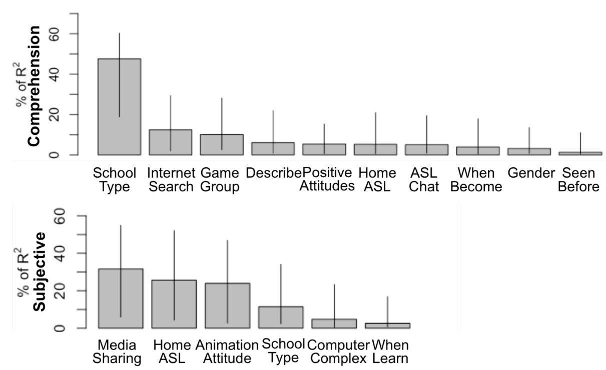

Figure 3: Relative importance (normalized to sum to 100%) of factors in Comprehension Model 2 and in Subjective Model 2, with 95% bootstrap confidence intervals.

ASL animation researchers generally assume that the value of such scores is based *upon the difference in quality or clarity of the animation stimuli that are shown to participants*. Perhaps counter-intuitively, the models in this paper do **not** include variables about the type of stimulus that was shown, e.g., the story nor the avatar, two variables that presumably relate to a participant's evaluation scores. Instead, we intentionally examined whether we could model the variance in scores based only upon demographic and experience/attitude characteristics of the participants.

7. DISCUSSION

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. In section 6, we considered each variable's coefficient ("Estimate" column in Tables 3 and 4) 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 [20], calculated using the 'relaimpo' package [7]. 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 3 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 3). 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 are SchoolType, InternetSearch, and GameGroup. Given the much higher relative importance of SchoolType, compared to the other variables, we focus on this variable in our discussion below:

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 for other studies (whose participant pools may have differed in this characteristic).

Comprehension and SeenBefore. Another aspect Figure 3 that may be of interest to sign language animation researchers is the low importance of the SeenBefore variable in this model. 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 re-recruited 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. We note that in this study, we had a well-balanced sample of participants 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 3 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 4.

Subjective and AnimationAttitude. A positive relationship exists 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 due to their possible enthusiasm for technology. On the contrary: MediaSharing 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.

Subjective and HomeASL. A participant using ASL at home was also a factor with a negative relationship to their subjective score. We speculate that this might also be a case of "higher standards"; frequent ASL users may be harsher critics of animation quality.

Subjective and SchoolType. While SchoolType was important in both Comprehension Model 2 and in Subjective Model 2, the *direction of the relationship is reversed*. 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 [13].

8. CONCLUSIONS AND FUTURE WORK

The long-term goal of our research is to improve the state of the art of software for automatically synthesizing animations of sign language from a simple script of the desired message, technology that would make it easier to maintain and update information online in the form of sign language. We are also interested in understanding how to best conduct studies to evaluate the quality of such software, and the findings of this current study will affect the set of demographic and technology experience/attitude questions we ask participants in future work. Thus, a contribution of this work 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 yes/no question.
- **MediaSharing:** Assessed with four scalar response items indicating frequency of different activities, from [25].
- **AnimationAttitude:** Assessed with six Likert agreement items.

While other variables were present in models presented in section 6, the above four items correspond to the most important factors in section 7, and this abbreviated set 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 will continue reporting other basic demographic data, e.g., age or gender, but our survey of prior work in 2.1 suggests that few current sign language animation researchers regularly collect and report these four items above.

We have previously released stimuli and evaluation questions to the research community, to promote replicability and comparison of results across studies [11]. We hope to further contribute by sharing the survey questions and ASL videos used in the study reported in this paper: http://latlab.ist.rit.edu/assets2015

Through collection and publishing of these characteristics of study participants, we anticipate easier comparisons of research results across publications. We also believe that these factors would 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, this study was relatively large (N=62). However, when conducting a regression analysis of factors, there is always an advantage in having even larger and more diverse participant sets. In this case, it would be useful to recruit more participants from the Deaf community in another geographic area, to ensure that the relationships observed in the current study are preserved.

In future work, we are also interested in further exploring the variable of Age. This variable was not selected by the exhaustive all-subsets model comparison in this study, but only 10% of our 62 participants were over age 43 (none ages 35-43). In future work, we would like to conduct additional targeted recruitment of older participants. As we have learned in this study, such participants were the most time-consuming to recruit; so, this must be factored into the data-collection timeline in future work.

9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under award number 1065009 and 1506786, and by the Swiss Science Foundation, under award 155263. We are grateful to Sandra Sidler-Miserez, who helped produce ASL animations, and to Kellie Menzies and Evans Seraphin, who assisted in the recruitment of participants and in data collection.

10. REFERENCES

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

[2] Crabb, M., Hanson, V.L. 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 & accessibility* (ASSETS '14). ACM, New York, NY, USA, 193-200. DOI=10.1145/2661334.2661356

[3] Ebling, S., Glauert, J. 2015. Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Univ Access Inf Soc.*

[4] Fox, J. and Monette, G. (1992) Generalized collinearity diagnostics. *JASA* 87:178-183.

[5] Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Stat Med* 27(15), 2865-2873.

[6] Gibet, S., Courty, N., Duarte, K., & Naour, T. L. 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), 23 pages. DOI=10.1145/2030365.2030371

[7] Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software* 17(1): 1-27.

[8] Hanke, T. 2001. *ViSiCAST Deliverable D5-1: Interface Definitions.* Technical Report. ViSiCAST project. http://www.visicast.co.uk/members/milestones/D5-1rev1.pdf

[9] Hayward, K., Adamo-Villani, N., Lestina, J. 2010. A computer animation system for creating deaf-accessible math and science curriculum materials. In *Eurographics'10.*

[10] Heloir. A, Nguyen, Q., Kipp, M. 2011. Signing avatars: a feasibility study. In *2nd Int'l Workshop on Sign Language Translation and Avatar Technology,* Dundee, Scotland.

[11] Huenerfauth, M., Kacorri, H. 2014. Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. *Workshop on the Representation & Processing of Sign Languages, LREC'14.*

[12] Huenerfauth, M., Kacorri, H. 2015. Best practices for conducting evaluations of sign language animation. *Proceedings of the 30th Annual International Technology and Persons with Disabilities Conference (CSUN'15), Scientific/Research Track,* San Diego, California, USA.

[13] Huenerfauth, M., Zhao, L., Gu, E., Allbeck, J. 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

[14] International Standards Organization. 2004. ISO/IEC 14496-2:2004: Information technology – Coding of audio-visual objects – Part 2: Visual.

[15] Jennings, V., Elliott, R., Kennaway, R., Glauert, J. 2010. Requirements For A Signing Avatar. *Workshop on the Representation & Processing of Sign Languages, LREC'10.*

[16] Kacorri, H., Huenerfauth, M. 2014. Implementation and evaluation of animation controls sufficient for conveying ASL facial expressions. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility (ASSETS '14).* ACM, New York, NY, USA, 261-262. DOI=10.1145/2661334.2661387

[17] Kacorri, H., Huenerfauth, M. 2015. Comparison of Finite-Repertoire and Data-Driven Facial Expressions for Sign Language Avatars. *In Proceedings of HCII2015.*

[18] Kennaway, J. R., Glauert, J. R., & Zwitserlood, I. 2007. Providing signed content on the Internet by synthesized animation. *ACM Trans Comput-Human Interact* 14(3): Article 15 (Sept 2007). DOI=10.1145/1279700.1279705

[19] Kipp, M., Nguyen, Q., Heloir, A., Matthes, S. 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, NY, USA, 107-114. DOI=10.1145/2049536.2049557

[20] Lindeman, R.H., Merenda, P.F., Gold, R.Z. 1980. *Introduction to bivariate & multivariate analysis*. Scott Foresman, Glenview, IL.

[21] Lumley, T. 2009. *leaps: regression subset selection. R package version 2.9.*

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

[23] Moemedi, K.A. 2010. Rendering an avatar from sign writing notation for sign language animation. Doctoral dissertation, University of the Western Cape.

[24] Prillwitz, S., Leven, R., Zienert, H., Hanke, T, Henning, J. 1989. *HamNoSys: v2.0: Hamburg Notation System for Sign Languages: an introductory guide.* Signum, Hamburg.

[25] Rosen, L.D., Whaling, K., Carrier, L.M., Cheever, N.A., Rokkum, J. 2013. The media and technology usage and attributes scale: An empirical investigation. *Comput Human Behav* 29(6): 2501-2511.

[26] Tran, J.J., Johnson, T.W., Kim, J., Rodriguez, R., Yin, S., Riskin, E.A., Ladner, R.E., Wobbrock, J.O. 2010. A web-based user survey for evaluating power saving strategies for deaf users of mobileASL. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '10).* ACM, New York, 115-122. DOI=10.1145/1878803.1878825

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

[28] VCom3D. 2015. Homepage. http://www.vcom3d.com

[29] Verlinden, M., Tijsseling, C., Frowein, H. 2001. Sign language on the WWW. In *Proceedings of 18th Int. Symposium on Human Factors in Telecommunication.*

[30] Yang, O., Morimoto, K., Kuwahara, N. 2014. Evaluation of Chinese Sign Language animation for mammography inspection of hearing-impaired people. In *Advanced Applied Informatics, IIAI'14,* 831-836. IEEE.