#

Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation

HERNISA KACORRI
PENGFEI LU
The City University of New York, Graduate Center
AND
MATT HUENERFAUTH
The City University of New York, Queens College

Many researchers internationally are studying how to synthesize computer animations of sign language; such animations have accessibility benefits for people who are deaf that have lower literacy in written languages. The field has not yet formed a consensus as to how to best conduct evaluations of the quality of sign language animations, and this article explores an important methodological issue for researchers conducting experimental studies with participants who are deaf. Traditionally, when evaluating an animation, some lower and upper baselines are shown for comparison during the study. For the upper baseline, some researchers use carefully produced animations, and others use videos of human signers. Specifically, this article investigates, in studies where signers view animations of sign language and are asked subjective and comprehension questions, whether participants differ in their subjective and comprehension responses when actual videos of human signers are shown during the study. Through three sets of experiments, we characterize how the Likert-scale subjective judgments of participants about sign language animations are negatively affected when they are also shown videos of human signers for comparison – especially when displayed side-by-side. We also identify a small positive effect on the comprehension of sign language animations when studies also contain videos of human signers. Our results enable direct comparison of previously published evaluations of sign language animations that used different types of upper baselines – video or animation. Our results also provide methodological guidance for researchers who are designing evaluation studies of sign language animation or designing experimental stimuli or questions for participants who are deaf.

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*.

This research was supported by grants from the National Science Foundation ("CAREER: Learning to Generate American Sign Language Animation through Motion-Capture and Participation of Native ASL Signers," Award #0746556, 2008; "Generating Accurate Understandable Sign Language Animations Based on Analysis of Human Signing," Award #1065009, 2011), from The City University of New York PSC-CUNY Research Award Program ("Educational Software for Deaf Users," 2008), from Siemens A&D UGS PLM Software ("Generating Animations of American Sign Language," Go PLM Grant Program, 2007), and from a free academic license for character animation software from Visage Technologies AB.

Authors' addresses: Matt Huenerfauth, Computer Science Department, Queens College/CUNY, 65-30 Kissena Blvd, Flushing, NY 11367 USA, telephone: 1-718-997-3264, email: matt@cs.qc.cuny.edu. Hernisa Kacorri, Pengfei Lu, Computer Science Program, CUNY Graduate Center, 365 Fifth Avenue, New York, NY 10016 USA, telephone: 1-212-817-8190, email: hkacorri@gc.cuny.edu; pengfei.lu@qc.cuny.edu.

General Terms: Design, Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Accessibility Technology for People who are Deaf, American Sign Language, Animation, Baseline, User Study.

ACM File Format:
KACORRI, H., LU, P., AND HUENERFAUTH, M. 2013. Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation. *ACM Trans. Accessible Computing*, #, #, Article # (Month 2013), XX pages. DOI = ##### http://doi.acm.org/######

1. INTRODUCTION

Because of limitations in language exposure and other educational factors, there are many deaf adults who have difficulty reading written-language text. For instance, in the U.S., a majority of secondary school graduates who are deaf (typically those who are age 18-21) have a fourth-grade (age 10) reading level or below [Traxler 2000]. So, while these adults can see the written text on television captioning, websites, or other media, they may not be able to access the information content, if the reading level of the text is too complex. For this reason, many accessibility researchers have begun investigating alternative techniques of presenting information to deaf users, including sign language animations. In the U.S., there are more than 500,000 people who use American Sign Language (ASL) as a primary means of communication [Mitchell et al. 2006]. Fluency in ASL and fluency in written English are distinct skills: ASL is a distinct natural language, with a different word order, syntax, and lexicon from English. (While there are a variety of different sign languages used internationally, the examples and discussion in this article focus primarily on ASL. However, the methodological issues explored in this article should be applicable to researchers studying other sign languages.) Because there are many deaf adults with more advanced fluency in ASL than in written English, providing information in the form of ASL can make more websites, computer software, and educational content accessible for these users [Huenerfauth and Hanson 2009].

While videos of human signers could be included in webpages or in computer software, there are many reasons why computer animations of sign language are more useful. For websites in which the information content is generated dynamically from a database, is frequently updated, or is customized for the user, it is impractical to provide videos of signing – because it would be prohibitively expensive to film a human performing ASL for the new/changing information. Technology for synthesizing ASL animations could also enable "scripting" and revision of messages – preserving the anonymity of the author (whose face would not be revealed as it would be in a video of ASL signing) or enabling the collaborative writing of ASL messages by multiple authors in a wiki-style setting. Because of the blending, modulation, and variability in how signing movements must appear to produce natural signing, it is not possible to achieve high-quality messages by stitching together a pastiche of videos of a human performing individual signs. Thus, researchers explore the synthesis of natural and understandable computer animations with virtual human characters performing sign language.

Our lab has conducted many studies with native ASL signers to evaluate the naturalness and understandability of ASL animations synthesized by our software [Huenerfauth and Lu 2012; Huenerfauth et al. 2011; Huenerfauth 2006]. Generally, we ask signers to watch ASL animations and then answer Likert-scale subjective questions and comprehension questions. We have also investigated methodological issues relating to the conduct of such studies, including: screening to identify native signers, designing ASL scripts that contain particular linguistic constructions, designing comprehension questions and answer sheets accessible to participants with low English literacy [Huenerfauth et al. 2008]. This article focuses on another important methodological issue: how the presentation of upper baseline for comparison (either a high-quality animation or a video of a human signer) affects the responses recorded in a study.

This article is an extended version of a paper originally presented at the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'12) [Lu and Kacorri 2012]. The studies in Section 4 were previously discussed in that paper. This article describes two new pairs of user studies (Sections 5 and 6) to investigate new research questions and to evaluate some hypotheses from our original paper that were not adequately evaluated due to limits in the size and design of our prior studies. This article also contains additional discussion of related literature (Section 2) and conclusions and advice for future researchers (Section 7).

1.1 Motivation of This Article

Like other research groups (discussed in Section 2), our lab investigates mathematical and linguistic models for synthesizing animations of sign language. To track our progress over time, we ask native signers to evaluate the understandability and naturalness of our animations. During these studies, we compare an animation that has been synthesized using our current model to some other animation (e.g., synthesized using an earlier model or using some simplistic "lower baseline" technique). Native signers answer subjective questions about the quality of our animations, and they answer comprehension questions about the animations' information content. A challenge when interpreting the results of comprehension questions is that the score is based on factors beyond the animation itself, e.g., a question's difficulty, a participant's memory skill, etc. To make it easier to interpret the results from comprehension questions, we have generally added an "upper baseline" (a third type of animation shown during the study for comparison). A good upper baseline should represent an "ideal" output of the system, and it may consist of a high-quality computer animation or a video recording of a human signer (performing identical sentences to the virtual human in the animations). As discussed in Section 2, research groups have differed in their selection of an upper baseline for evaluations: some have used videos of humans, and some have used computer animations. There are trade-offs for either choice, as discussed below.

As an upper baseline in our past studies, we generally used a computer animation of a virtual character that is visually identical in appearance to the virtual human in our model-synthesized animations. We ask a native ASL signer who is a skilled animator to carefully control the movements of the character to produce the most fluent/natural animation possible – performing identical sentences to the virtual human in our model-synthesized animations. The rationale for this choice was that our laboratory does not investigate issues related to the photorealism of virtual human animations, but instead, we investigate models of the movements of the virtual human character. Thus, an animator-controlled high-quality animation represents an "ideal" output of what our software could achieve. Further, we were wary of including videos of real humans in our experiments, because we were concerned that participants would focus on the differences in appearance between the human and the virtual human – and thereby attend less to the movement subtleties of the virtual human character, which was our research focus.

On the other hand, an intuitive upper baseline for an experiment would be a video of a human ASL signer. Our study participants are used to seeing humans performing ASL all the time; they are less familiar with seeing computer animations of ASL. Further, a video of a native ASL signer performing ASL would likely have higher fluency/clarity than any animation; so, it could be considered a truer "ideal." Another advantage is that non-experts can interpret the results from the experiment; since a video of a human is more familiar than an animation of a virtual human signing, it is easier to understand the results of the experiment, relative to a human video upper baseline. A downside of a human video upper baseline is that it may be an impossibly high ideal – the state of the

art of sign language computer animation may be decades away from producing something with similar quality to a video of an actual human. So, a video upper baseline could yield scores that are so higher "off the scale" that they could make it difficult to obtain meaningful evaluation scores for the other animations shown in the study.

Despite these various trade-offs and despite prior research groups making different choices as to their upper baselines (details in Section 2), we have not found any prior methodological research on the effect of selecting each of these different types of upper baseline. While it is intuitively plausible that a computer animation being evaluated may receive different evaluation scores in an experiment – depending on whether it is being compared to another animation or compared to a video of a human, the specific empirical effect of selecting an upper baseline has not been quantified. This means that currently there is no reliable way to compare the empirical results across evaluation studies conducted by different research groups (who have used different upper baselines), and there is no guidance for future researchers as to the best approach to use when designing their evaluation studies. The goal of this article is to fill this gap in the methodological literature and to provide a useful foundation for future empirical research in the growing field of sign language computer animation synthesis.

1.2 Overview of This Article

This article begins with a survey of related work on conducting evaluations of computer animations of sign language (or virtual humans performing other actions) with a focus on the types of upper baselines used in those studies (Section 2). Next, a set of research hypotheses are outlined that relate to how the selection of a video or animation upper baseline affects the results of a user study (Section 3). Section 4 describes our first experimental study, in which we replicate a prior study [Huenerfauth and Lu 2010] and replace the upper baseline in that study with a video of a human signer. Section 5 describes a previously unpublished pair of experiments, focused on facial expressions during sign language, with a similar structure: performed once with an animation upper baseline and then performed again with a video upper baseline. Section 6 presents a third set of experiments that explores a related issue: whether presenting the comprehension questions in an experiment in the form of animation or video affects the evaluation results. Finally, Section 7 presents our conclusions, advice for future researchers, and future work.

2. RELATED WORK

Many researchers have studied the usability of computer animations of virtual humans in various applications, including comparisons to videos of humans, e.g. [Ham et al. 2005; McDonnell et al. 2008; Russell et al. 2009]. We searched the literature for examples of prior user studies (of both sign language animation and non-signing virtual human animation), and we noted the type of upper baseline used in those studies. We found that prior user studies can be organized into three categories.

The first category is research in which no upper baseline is mentioned. Although evaluation against a baseline usually results in more meaningful scores, many user-studies don't include any baselines. For example, researchers asked users to evaluate their deployed human animation without any baselines for comparison [Schnepp et al. 2010], improved their animations through iterative experiments [Davidson et al. 2000], defined the best parameters for their animation models through presentation of multiple versions [Davidson et al. 2000], compared the suitability of their available animated characters for a given task [Ow 2009], or compared their results to previous similar studies [López-Colino et al. 2011].

The second category is where video recordings of a human are used as an upper baseline for comparison to the animation under evaluation. For example, researchers have assessed comprehensibility while comparing their avatars to human signers [Kipp et al. 2001a], verified the visual quality of their animations by comparing them to video of human interpreters [Baldassarri et al. 2009; Baldassarri and Cerezo 2012], or compared interpreter videos to animations in a medical domain [Morimoto et al. 2003; Morimoto et al. 2006]. Researchers studying non-signing virtual characters have also sometimes used videos of humans as a baseline for comparison. For example, the expressiveness of a MPEG-4 face model [Ahlberg et al. 2002] and eye gazing in humanoid avatars in dyadic conversations [Garau et al. 2001] have been evaluated in comparison to human videos.

The last category is research in which animations of a virtual character were used as an upper baseline in the evaluation; this seems to be the most popular approach for sign language animation (and non-signing virtual human or embodied agents) research. We noticed that the similarity of appearance between the virtual characters in the "upper baseline" animation and the character in the animation under evaluation varied across studies, and so, we decided to divide this research into two subcategories, according to the way in which the upper baseline animation was created and manipulated.

The first subcategory of prior research used upper baseline animations that were controlled by a human animator, without any motion-capture data. This is the approach we have used in prior studies at our lab, in which we asked a human animator to carefully produce an animation of a virtual human, with the same appearance as the animation for evaluation, to serve as our upper baseline [Huenerfauth et al. 2011; Lu and Huenerfauth 2011; Lu and Huenerfauth 2012]. Researchers studying the animation of non-signing virtual human characters have employed a similar methodology, e.g. [Bergmann 2012] compared "average" models learned from the combined data of several speakers with individualized generated gestures based on empirically observed gestural behavior.

The second subcategory includes studies where the upper baseline was an animation produced, at least partially, from a motion-capture recording of a human. Sign language animation researchers have used this type of upper baseline in a variety of studies: to rate the understandability, naturalness of movement, and grammaticality of animations [Huenerfauth 2006]; to measure the comprehension of synthesized facial expressions [Gibet et al. 2011]; to evaluate synthesized signs [Kennaway et al. 2007]; or to elicit feedback on a variety of signing animations [Kipp et al. 2011b]. These papers are listed in Table 1, which highlights the degree to which the upper baseline animation was similar to the other animation being evaluated in the study. When an "X" appears in a column, it means that the upper baseline shared a property with the animation being evaluated: the language of signing, the content of the signed message, the animation tool used to produce the animation, the appearance of the virtual human, and the background of the animation. In addition, researchers studying non-signing animations have also used virtual humans driven by motion-capture as upper baselines [Pražák et al. 2010].

Table 1: Similarity of the upper baseline animation to the animation being evaluated.

	Gibet et al. 2011	Huenerfauth 2006	Kennaway et al. 2007	Kipp et al. 2011b
Language	x	x	x	-
Message Content	x	x	x	-
Animation Tool	x	x	-	-
Character Appearance	x	-	-	-
Animation Background	x	x	x	-

During our review of the literature, we also noted how instructions or comprehension/evaluation questions were displayed to participants. Based on the modality of presentation, we identified four categories of prior studies (listed below). Section 6 will investigate how the modality of presentation of study elements beyond the stories themselves (i.e., the comprehension questions) may affect the results.

- The first category includes studies in which a *human experimenter* signs: instructions [Schnepp et al. 2010; Kipp et al. 2011a; Kennaway et al. 2007], questions [Davidson et al. 2000], or guidance to a focus group [Kipp et al. 2011b].
- The second category includes studies where *video recordings* of a human are used to present instructions within the user-interface of the software displaying the animation stimuli [Schnepp and Shiver 2011; Schnepp et al. 2011] or provided as explanations for the questions in an online study [Kipp et al. 2011b].
- The third category includes studies in which an *animated character* (similar in appearance to the virtual human in the animations being evaluated) performs comprehension questions [Huenerfauth and Lu 2010; Lu and Kacorri 2012].
- The last category includes studies in which *written text* is used to present study instructions [Baldassarri et al. 2009; López-Colino and Colás 2011] or questionnaires [Gibet et al. 2011; Ow 2009].

3. RESEARCH METHODOLOGY & HYPOTHESES

To compare the results of prior studies that used different upper baselines (Section 2), we need to quantify how the upper baseline affects the evaluation scores collected. To do so, we need to conduct an identical study in two ways: (1) once using videos of human signers as an upper baseline and (2) once using computer animations as an upper baseline. If the other animations shown in the study (aside from the upper baseline) remain constant, then any differences in their evaluation scores could be attributed to difference in the upper baseline used. In this manner, we can examine several research questions:

- Do video upper baselines receive higher comprehension question scores than animation upper baselines do?
- If a study uses a video upper baseline (instead of an animation upper baseline), then are the comprehension scores for the other animations in the study affected?
- Do video upper baselines receive higher Likert-scale subjective evaluation scores than animation upper baselines do?
- If a study uses a video upper baseline (instead of an animation upper baseline), then are the Likert-scale subjective evaluation scores for the other animations affected?

It is important to note that, for this article, the scientific aim of any individual study (determining if the mathematical/linguistic model under consideration produces good ASL animations) is not important: Instead, we are only focused on whether changing the upper baseline in the experiment causes measurable differences in the evaluation scores for the upper baseline and for the other animations being evaluated in that experiment.

In order to formulate some hypotheses in regard to these questions, we considered a pair of prior experiments at our lab that were nearly (but not exactly) identical: with one experiment using an animation upper baseline and the other using a video upper baseline. In [Lu and Huenerfauth 2010], we conducted an experiment to evaluate the quality of some computer animations of sign language, and we used an animation produced by a

human animator as an upper baseline. In [Lu and Huenerfauth 2012], we conducted a study to evaluate a similar (but not identical) set of computer animations of sign language, but we used a video of a human signer as an upper baseline. In both studies, native ASL signers who saw the animations/videos answered comprehension questions and Likert-scale subjective evaluation questions. Unfortunately, this prior pair of studies was not a perfect test of our research questions: The script of the ASL stories in the two studies was not identical (so the stories might have been harder in one of the studies). Further, the human in the videos used as an upper baseline in [Lu and Huenerfauth 2012] wore some motion-capture equipment; so, he may have been harder to understand. Regardless, by considering how the scores in these (approximately) identical studies differed, we gain insight into the effect of different upper baseline – and can formulate some hypotheses.

Changing the upper baseline did not produce a difference in the comprehension question scores for the other stimuli in the study (the motion-capture-based animations), which had similar scores in both studies. For the Likert-scale subjective scores (1-to-10 scales for naturalness of movement, perception of understandability, and grammatical correctness of the animations), in the later study (with the video upper baseline), the other animations received lower scores than they had in the prior study. We speculate that seeing a video of a human as one of the stimuli led participants to assign lower Likert-scale subjective scores to the animations (which looked worse by comparison to a video of a real human). Based on these prior studies, we hypothesize the following:

- H1: A human video upper baseline will receive higher comprehension question scores than an animated-character upper baseline produced by a human animator.
- H2: The upper baseline used (human video or animated character) will not affect the comprehension questions accuracy scores for the other stimuli shown in the study.
- H3: A human video upper baseline will receive higher Likert-scale subjective scores than an animated-character upper baseline.
- H4: Using a human video upper baseline will depress the subjective Likert-scale scores that participants assign to the other stimuli in the study.

As mentioned at the end of Section 2, videos of human signers could appear during a user study in other capacities, aside from appearing as an upper baseline. Specifically, videos of human signers might be displayed to study participants as the comprehension questions that participants are asked after viewing each of the sign language animations. Displaying videos of humans asking questions in ASL could also have an effect on participants' subjective ratings of the animations in the study. Therefore, Section 6 will evaluate the following two additional hypotheses, which focus on a comparison between presenting questions as videos of human signers or as animations of sign language:

- H5: Displaying comprehension questions as videos of a human signer or as a high-quality animation will not affect the comprehension questions accuracy scores.
- H6: Displaying comprehension questions as videos of a human signer or as high-quality animations will not affect the subjective Likert-scale scores that participants assign to the animations in the study.

3.1 Three Phases of This Research

Our research methodology consists of three phases, which are summarized in Table 2. Portions of the work in Phase 1 were described previously in [Lu and Kacorri 2012].

Table 2: Summary of Three Phases of Experiment

			Upper Baseline	Model	Lower Baseline	Comprehension Questions		
Effects of Video Upper Baseline Hypotheses: H1 - H4	PHASE 1 Hand Movements	2010 18 participants	Animation verb created by a native signer	Animation verb sinthesized by our model	Animation dictionary-entry version of the verb	Animation	Part 1 9 stories	Part 2 8 stories
		2012 18 participants	Human Video				Part 1 9 stories	Part 2 8 stories
	PHASE 2 Facial Expressions	2013 16 participants	Animation facial expressions defined by a native signer	Animation facial expressions defined by our model	Animation without facial expressions	Human Video	Part 1 21 stories	Part 2 7 stories
		2013 18 participants	Human Video				Part 1 21 stories	Part 2 7 stories
Effects of Video Questions Hypotheses: H5 - H6	PHASE 3 Hand Movements	2010 18 participants	Animation verb created by a native signer	Animation verb synthesized by our model	Animation dictionary-entry version of the verb	Animation	Part 1 9 stories	N/A
		2013 18 participants				Human Video	Part 1 9 stories	N/A

In Phase 1 (section 4 of this article), to evaluate hypotheses H1-H4, we conducted an identical pair of experiments, with the only difference being the upper baseline used. Participants evaluated animations of short stories that contained ASL verbs with complex hand movements. Section 4 portrays the challenges in recording videos of a human performing an identical script of signs as an animated character. The clearest results were for hypotheses H3 and H4. The video upper baseline received higher Likert-scale subjective scores (H3 supported), and it led to lower Likert-scale subjective scores for the other stimuli during the side-by-side comparison part of the study (H4 supported).

In Phase 2 (section 5), an additional pair of studies was conducted: one with a video upper baseline and one with an animation upper baseline. Our lab recently began studying the synthesis of facial expressions for sign language animations; so, for Phase 2, we used animations of ASL sentences with various grammatical and emotional facial expressions. The studies in Phase 2 also contained a larger number of comprehension questions, to enable us to better evaluate hypotheses H1 and H2, which were not adequately addressed in Phase 1. The video upper baseline received higher comprehension scores than the animation upper baseline (H1 supported) and led to a small increase in the comprehension scores for the other stimuli (H2 not supported).

In Phase 3, to evaluate hypotheses H5 and H6, we conducted a final pair of studies: one with comprehension questions presented as human videos and one with comprehension questions presented as high-quality ASL animations. Section 6 describes how the choice of video or high-quality animation had no effect on the comprehension (H5 supported) or Likert-scale scores we collected (H6 supported).

4. PHASE 1: ANIMATION VS. VIDEO UPPER BASELINES

To clearly evaluate hypotheses H1-H4, we needed to compare the results of two experiments that were identical, aside from the upper baseline used. Since we had previously conducted a study in which computer animations were used an upper baseline [Huenerfauth and Lu 2010], we decided to replicate that study. We replaced the upper baseline with a video of human signer performing identical ASL stories as the animated character. This section describes the challenges we faced when producing a video of a

human that performed identical signs to our animated character upper baseline, and it presents the results of our 2012 replication of our original 2010 study.

4.1 Design of Evaluation Studies for Phase 1

In [Huenerfauth and Lu 2010], we designed a model for synthesizing the movements of "inflected" ASL verb signs whose movements depend on locations in the space around the signer where the verb's subject and object have been previously set up. In order to evaluate the understandability of animations in which the verbs were produced using our new model, we compared three versions of animations: (1) a lower baseline consisting of the simple dictionary-entry versions of the verb signs (where the hand movement doesn't indicate subject/object), (2) a middle case consisting of animations of the verbs synthesized by our model, and (3) an upper baseline consisting of animations of inflected versions of each verb produced by a human animator, who was a native ASL signer.

The experimental study consisted of two parts: In part 1, participants were asked to view 9 animations of a virtual human character telling a short story in ASL. Each story included instances of the inflected verbs. A fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 3 animations of each version: i) lower baseline, ii) model, or iii) upper baseline. Fig. 1 shows a story transcript, the English translation for this transcript is "Hi, my name is Charlie. I have three friends. Bob, Sue and Jason. Jason has an old book from library, he gives the book to Sue. The book is due tomorrow and it must go to library. Sue asks Bob where the library is. Bob doesn't know. Bob asks Jason where the library is. Jason tells Bob the library is on 9th street. Sue tells Jason the library is closed. She gives the book to Bob. Tomorrow Bob will go to the library." Colors indicate locations around the signer where the verb's subject/object are located. After watching each story animation (of one of three types: lower baseline, model-synthesized, or upper baseline) one time, participants answered 4 multiple-choice comprehension questions. Questions focused on whether they understood and remembered the subject and object of each verb. Participants also responded to three 1-to-10 Likert-scale questions about how grammatically correct, easy to understand, or naturally moving the animation appeared.

HI. MY NAME #CHARLIE. I HAVE THREE FRIENDS.

#BOB THERE$_{PURPLE}$. #SUE THERE$_{BLUE}$. #JASON THERE$_{ORANGE}$.

HE$_{ORANGE}$ HAVE OLD BOOK FROM LIBRARY. BOOK HE$_{ORANGE}$ GIVE$_{ORANGE \rightarrow BLUE}$ HER$_{BLUE}$.

TOMORROW BOOK DUE MUST GO LIBRARY.

SHE$_{BLUE}$ ASK$_{BLUE \rightarrow PURPLE}$ HIM$_{PURPLE}$ WHERE LIBRARY. HE$_{PURPLE}$ DON'T KNOW.

HE$_{PURPLE}$ ASK$_{PURPLE \rightarrow ORANGE}$ WHERE LIBRARY.

HE$_{ORANGE}$ TELL$_{PURPLE}$ HIM$_{PURPLE}$ LIBRARY ON 9TH STREET. SHE$_{BLUE}$ TELL$_{ORANGE}$ HIM$_{ORANGE}$ LIBRARY CLOSED.

BOOK SHE$_{BLUE}$ GIVE$_{BLUE \rightarrow}$ HIM$_{PURPLE}$. TOMORROW HE$_{PURPLE}$ GO LIBRARY.

Fig. 1. Example script for a story shown in the study.

In part 2 of the study, we used a side-by-side comparison methodology described in detail in [Huenerfauth and Lu 2012; Huenerfauth et al. 2008]: participants viewed three versions of an animation of a single ASL story side-by-side on one screen, as depicted in Fig. 2(a). A total of 8 stories (three versions of each) were viewed by each participant. The sentences shown side-by-side were identical, except for the version of the verb that appeared in each, which was either: the dictionary-entry version of the verb animation

(the "lower baseline"), the verb animation synthesized by our model, or the verb carefully created by a native ASL signer using animation software [Vcom3D 2012] ("upper baseline"). The participants could re-play each animation as many times as they wished. Participants were asked to focus on the verb and respond to a single 1-to-10 Likert-scale question about the quality of the sentence animation. (We note that "part 2" of the study always followed "part 1" of the study, and this could be seen as a limitation in the study design. In future work, we may counterbalance the order of these parts.)

Fig. 2. Screenshots from the side-by-side comparison, as seen by participants in (a) 2010 or (b) 2012.

During our replication of the study, in 2012, we replaced the upper baseline animations with videos of a human signer. The top row in Fig. 2(a) shows an example of what the participants saw in the part 2 (side-by-side comparison) in 2010 and the lower row in Fig. 2(b) shows what they saw in 2012. All the other animations and their sequencing in this pair of studies were identical.

4.2 Recording the Human Video Upper Baseline for Phase 1

To produce the human video upper baseline, we recorded a native signer in our studio. For part 1, we needed to record a human performing the 9 short stories (with inflected versions of the verbs), and for part 2, we recorded a human performing the 12 sentences (with inflected versions of the verbs). Producing a video recording of a human that "matched" the animations being shown in the study was challenging. We wanted to "control" as many of the variables of the ASL performance between the upper baseline and the animation under primary evaluation (our model-synthesized animation) as possible, so that it would serve as an effective upper baseline for comparison. To match the background color, the human sat in front of a blue curtain. The human signer also wore a green t-shirt on the day of the recording, which was similar to the virtual human. To maintain the same viewing perspective, we placed the camcorder facing the signer at his head height, which matched the "virtual camera position" in the ASL animations. We cropped and resized the video files to match the height/width of the 2010 upper baseline animations – and to approximate the same placement of a human in a the video frame as

how the virtual human character had appeared in the animation in 2010. The framerate and the resolution of the video were identical to the animation from 2010.

To ensure that the human signer performed fluent ASL signing, all of the instructions and interactions for the recording session were conducted in ASL by another native signer sitting behind the camcorder. We needed the human signer to perform the same "script" of signs as the other (animation) stimuli shown in the study; so, we placed a large monitor in front of the signer (just below the camera) to display the story scripts (a script example is shown in Fig. 1). The signer had time to memorize and practice each of the scripts prior to the recording session – so that he would not need to read the script while signing. Because the stories were a bit complicated (an average of 55 signs in length, included 3-5 main characters set up at various locations in the signing space, with 3-5 inflected verbs per story), the signer had to practice in order to perform each story fluently. Unfortunately, due to the complexity of the stories and the need for accuracy, the signer found it very difficult to avoid glancing at the script occasionally during the performance. While we asked the signer to attempt to memorize the script and not look at the script during the recording process, several of our recordings contained infelicitous moments when the signer's eyes glance at the monitor displaying the story script.

On one hand, we wanted to control as many variables as possible so that they were held constant between our upper-baseline video and our animation being evaluated; on the other hand, we wanted to record a natural, fluent version of the sentences from the human signer. If the human's performance looks artificial, then it would not be an ideal of fluency and naturalness. Since ASL has no standard written form, and multiple signs can have similar meaning and use the same notation used in our notation, we had to explain our notation scheme to the signer and occasionally demonstrate which ASL sign was indicated by a particular word in the script. The script notation does not capture all of the subtleties of performance that are part of ASL; it is merely a loose sketch of what must be signed. While there was a script, the entire ASL performance was still underspecified, leaving room for the human, who was a native signer, to fill in the remaining elements of the performance based on his linguistic expertise in ASL.

While we gave the signer some "artistic freedom" in the performance of the stories, for the sake of naturalness and fluency, we did have to ask the signer to control the speed of his signing and some aspects of facial expressions, torso movement, head movement, etc., that could not be supported by the current animation tool and were not included in the animations being evaluated. To produce the same time duration in the videos as the upper baseline animations that had been used in 2010, we asked the signer to practice several times before the recording, and we used a stopwatch to measure how many seconds he took for each story during the practice and recording. After making several recordings of each story, we picked the one video recording of the story with the closest time duration to the upper baseline animation from 2010. We also asked the signer not to add too many theatrical embellishments, e.g., additional emotional facial expressions, which hadn't appeared on our virtual human character's face. This coaching and scripting process that was required in order to produce a good-quality human video upper baseline was surprisingly time-consuming, and it often felt like a delicate "balancing act" between guiding/controlling the human's performance while still allowing freedom in the performance so that the result would be natural and fluent. Even with this complex process outlined above, our resulting videos may not have been completely natural and fluent. For instance, some of the participants noticed problems in the human video, e.g., commenting that the "person signs well but need[s] little [more] facial expression."

One aspect of ASL that is especially difficult to capture in a script notation is specifying where in the space around a signer's body someone should point to refer to entities under discussion. Because the stories in this experiment contained many characters or objects that were assigned locations in space, with some verb signs in the stories changing their hand movements based on these locations, we needed our human performer to point to particular locations in space during the story (that identically matched the locations where our virtual human character points in the animations). Fig. 3 illustrates how we set up small colored paper squares around the studio (with colors that matched the script in Fig. 1) to guide the human where to point or where to aim the motion path of inflecting verb signs during the recording session. At 30-degree intervals around the signer, the squares were arranged in an arc in the following order (from the signer's left to the signer's right): purple, white, red, green, blue, orange, and yellow.

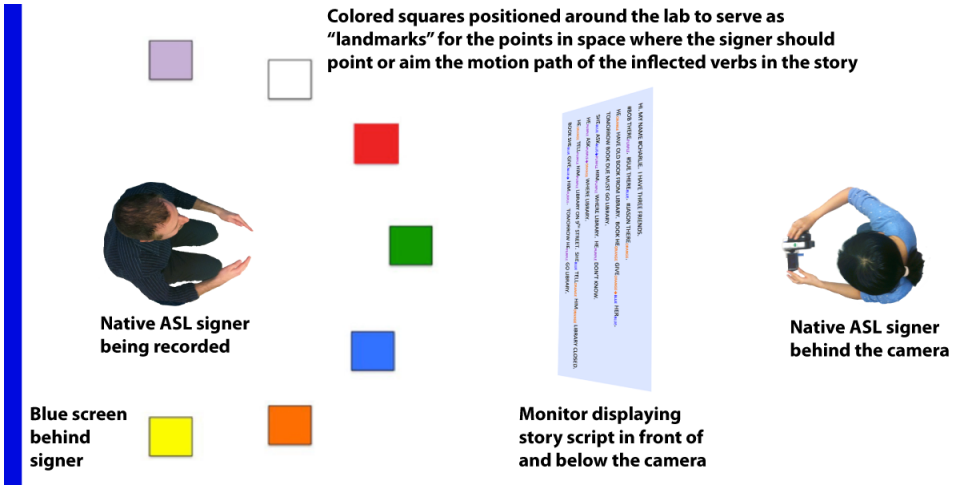

Fig. 3. Diagram of an overhead view of recording studio.

4.3 Data Collection and Results for Phase 1

We note that, given the goal of this study, it is not possible to test our hypotheses with a fully within-subjects design. Once a participant has seen an upper baseline video of a human signer, then they cannot participate in the animation upper baseline portion of the study. There is a carry-over effect: the participants cannot "un-see" or forget the video upper baseline once it has been seen. Further, there may be a practice effect when viewing animations of ASL and answer comprehension questions, and since it would not be possible to counterbalance the order in which participants participate in each study, we could not control for this order effect. Fortunately, we were able to design the experiments to control some variability due to individual differences in participants' skill. Identical recruitment procedures were followed in both 2010 and 2012, and very similar demographics were observed in the participants in both studies.

To ensure that responses given by participants are as linguistically accurate as possible, our lab screens participants to ensure that they are native ASL signers and controls the experimental environment so that it is ASL-focused; details of these methods appear in [Huenerfauth et al. 2008]. A native-signer conducted all of the interactions for our studies. Ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL-

based school as a young child. The 2010 study included 18 participants: 12 learned ASL prior to age 5, and 4 attended residential schools using ASL since early childhood. The remaining 2 participants used ASL for over 15 years, learned ASL as adolescents, attended a university with instruction in ASL, and used ASL daily to communicate with a family member. There were 12 men and 6 women of ages 20-56 (average age 30.5). The 2012 study included 18 participants: 16 learned ASL prior to age 5, and 10 attended residential schools using ASL since early childhood. The remaining 2 participants used ASL for over 13 years, learned ASL as adolescents, attended a university with instruction in ASL, and used ASL on a daily basis to communicate with a family member. There were 12 men and 6 women of ages 22-49 (average age 32.8).

The results collected include the comprehension-question and Likert-scale scores in part 1 of the studies (after a participant viewed a story one time) and the Likert-scale scores collected in part 2 of the studies (in which participants assigned a score to each of the three sentences which they viewed side-by-side). In Fig. 4, 5, and 6, which display the results, the thin error bars display the standard error of the mean. Animator10 and Video12 were the upper baselines, Lower10 and Lower12 were the lower baselines with the dictionary-entry version of the verbs, and Model10 and Model12 were the versions of the animations produced using our verb model. It is important to note that Lower10 and Lower12 were identical stimuli; the only difference was that the evaluation scores were collected in either the 2010 or 2012 study – likewise for Model10 and Model12.

To evaluate some of our hypotheses, we needed to consider the union of the responses for the Model and Lower animations in each study; these are displayed as "Model+Lower10" and "Model+Lower12" in Fig. 4, 5, and 6. Thus, "Model+Lower10" includes all of the data for Lower10 and Model10 combined. For the sake of clarity, we have included two graphs in each figure so that we would never display "Model+Lower" in the same graph as "Model" or "Lower" (since the latter is the combination of the data of the former two). In each figure, the graph on the left includes data for the upper baselines and for the "Model+Lower" data, and the graph on the right includes the individual results for Lower, Model, and the upper baselines.

One-way ANOVAs were used for comprehension-question data to check for statistical significance, and Kruskal-Wallis tests, for Likert-scale scores (because the Likert-scale data was not normally distributed). Statistical significance ($p<0.05$) for any of our planned comparisons has been marked with a star in Fig. 4, 5, and 6. The following comparisons were planned and conducted: (1) all three values from 2010, (2) all three values from 2012, (3) Video12 and Animator10, (4) Model12 and Model10, (5) Lower12 and Lower10, (6) Model+Lower10 and Model+Lower12, (7) Animator10 and Model+Lower10, and (8) Video12 and Model+Lower12. The reader should note that comparisons (1), (2), (7), and (8) are not needed to evaluate the specific hypotheses in this article. A researcher evaluating the quality of an animation relative to upper and lower baselines would traditionally perform these comparisons, and so we have presented them here for the benefit of future researchers who want to compare their results to ours.

Since H2, H5, and H6 are null hypotheses, it is appropriate to conduct "equivalence testing" to determine if pairs of values are indeed statistically equivalent, and we have therefore followed the two one-sided test (TOST) procedure [Schuirmann 1987], which consists of: (1) selecting an equivalence margin *theta*, (2) calculating appropriate confidence intervals from the observed data, and (3) determining whether the entire confidence interval falls within the interval (-*theta*, +*theta*). If it falls within this interval, then the two values can be deemed equivalent. We've noticed that in our prior work, when we found significant differences between groups of sign language animation, we've

generally seen differences of at least 1.5 Likert-scale units or 15% comprehension-question accuracy scores (e.g., [Huenerfauth and Lu 2012]). Thus, we've selected equivalence margin intervals of (-1.5, +1.5) for Likert scores and (-0.15, +0.15) for comprehension scores. Having selected an alpha-value of 0.05, then according to the TOST procedure for a two-sided analysis, we use a 90% confidence interval. Equivalence testing has been performed for the following pairs of values: (a) Video12 and Animator10, (b) Model12 and Model10, (c) Lower12 and Lower10, and (d) Model+Lower10 and Model+Lower12. Confidence intervals were determined using t-tests (for comprehension-question data) or Mann-Whitney U-tests (for Likert-scale data).

The data analysis and creation of the graphs for Phase 2 and Phase 3 have been conducted in an analogous manner for those studies, and therefore the above details are not repeated again in Sections 5 and 6 of this article.

Fig. 4 illustrates the comprehension-question accuracy scores from the 2010 and 2012 studies. Hypothesis H1 would predict that Video12 would have significantly higher scores than Animator10, but this was not supported by the data. This was an interesting result: our videos of a human signer did not achieve higher comprehension scores than the upper baseline animations we produced in 2010 of a virtual human with the verbs carefully planned by a human animator. We speculate that our challenges in recording the human video may have led to some understandability problems in Video12 stimuli, or this may indicate that our upper baseline animations from 2010 were of good quality.

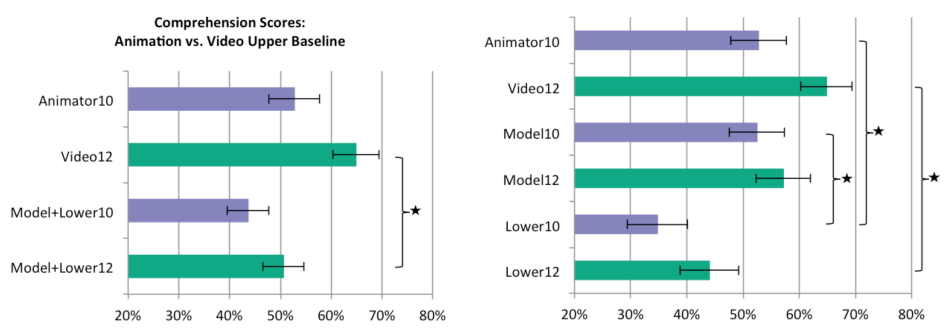

Fig. 4. Results of comprehension scores in Phase 1.

Hypothesis H2 would predict that the comprehension scores for the Model and Lower stimuli would be unaffected by changing the upper baseline from an animation in 2010 to a video in 2012. The following confidence intervals were calculated for TOST equivalence testing: (-0.154, 0.014) for Model+Lower10 vs. Model+Lower12, (-0.160, 0.067) for Model10 vs. Model12, and (-0.216, 0.031) for Lower10 vs. Lower12. Given that these intervals are not entirely within our equivalence margin interval of (-0.15, +0.15), we cannot determine whether pairs are equivalent. Thus, H2 is inconclusive. In Phase 2, we will conduct a study with more response data to better investigate H2.

Fig. 5 illustrates the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness that participants answered after watching the short stories in the studies. Hypothesis H3 was supported by the data in Fig. 5; the video upper baseline received higher subjective Likert-scale scores than the animation upper baseline. All three scores had the same pattern: Video12 had significantly higher grammaticality, understandability, and naturalness scores than Animator10.

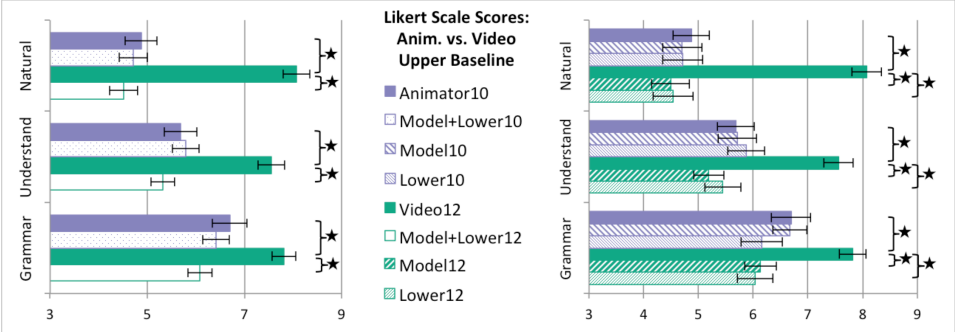

Fig. 5. Results of grammaticality, understandability, and naturalness Likert-scale scores in Phase 1.

Hypothesis H4 was that the use of a video upper baseline would lead to a change in the Likert-scale subjective scores for the other stimuli in the study. The data in Fig. 5 did not support hypothesis H4; there was no significant change in the Likert-scale scores for Model or Lower when we used the video upper baseline in 2012. When we examine the Likert-scale scores obtained during side-by-side comparisons in Fig. 6, we will see some contradictory results in regard to Hypothesis H4.

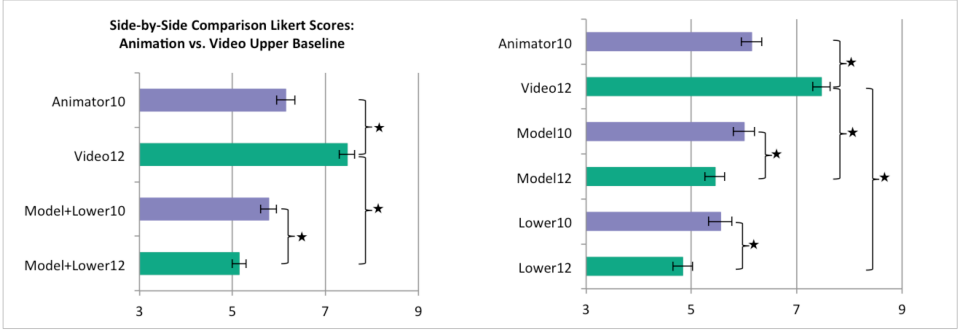

Fig. 6. Results of side-by-side comparison scores in Phase 1

Fig. 6 illustrates the Likert-scale subjective scores collected from participants during part 2 of the studies (the side-by-side comparison of identical sentences, with different versions of the verb in each, which could be replayed many times). Video12 is significantly higher than Animator10, further supporting Hypothesis H3 that video upper baselines would get higher Likert-scale subjective scores than animation upper baselines.

The results in Fig. 6 supported Hypothesis H4: Using a human video upper baseline depressed the subjective Likert-scale scores that participants gave to the animations. The 2012 values for Model+Lower, Model, and Lower were all significantly lower than their 2010 counterparts. The magnitude of this depression is 10%-20%. This is not a surprising result; when looking at videos of humans in direct comparison to animations of a virtual human character, it is reasonable that participants would feel that the animations are lower quality. What is surprising is that we had not observed any significant depression in Fig. 5 when looking at the Likert-scale data from part 1, in which participants assigned a Likert-scale subjective score to a story that they had just watched. We speculate that the depressive effect may depend on whether participants are assigning Likert-scale subjective scores to videos in a side-by-side direct comparison (as in part 2, Fig. 6) or sequentially throughout a study (as in part 1, Fig. 5). Perhaps the side-by-side setting forced participants to be more comparative – with the video "standing out" from the other

two stimuli, which were both animations. Another possible explanation for this result may be that during part 1 of the study, when watching a story one time and then answering the comprehension questions, the participants may have been very focused on the task of trying to understand and remember as much information as possible from the stories. Thus, they may have been less focused subjectively on the superficial appearance of the animations/videos. We will explore H4 further in Phase 2 in this article.

As discussed in Section 4.2, when creating baselines for comparison to animations in a study, a balance must be achieved between matching the content of the stimuli across versions and allowing for natural signing. Some of the comments of participants in the study indicated that in a few cases, we were not successful at this. Specifically, when producing the script for the human to perform in the video recordings, we included every sign that was performed by the virtual human character in the upper baseline animations from 2010. When a signer sets up points in space to represent entities under discussion, the signer may refer to these items later in the conversation by pointing to them. Because the movement path of an inflected ASL verb indicates the location around the signer where the subject and object of the verb are established, it is common (but not required) for signers to omit pointing to the subject/object before/after the verb (because the location in space that represents those entities is already indicated by the motion-path of the verb). The human animator who produced our upper baseline animations in 2010 still included some extra "pointing" to these locations, and so we included them in the script given to the human signer in 2012. In the feedback comments in 2012, some participants said: "Most verbs shouldn't end with the pointing of the finger (or direction) as the action already indicated that much," "too many endings were a pointing, it threw off my attention a lot," etc. What is interesting is that no participants criticized this in 2010; thus, when they saw a human signer performing this extra pointing movement, it felt more unnatural and warranted a comment at the end of the study.

5. PHASE 2: ANIMATION VS. VIDEO BASELINES WITH FACIAL EXPRESSION

A trend in the sign language animation literature is toward investigating facial expressions, and our lab intends to pursue this line of research in future work. Thus, for Phase 2 of this article, we decided to conduct another round of experiments with ASL animations with facial expressions. We also wanted to conduct a study with a larger number of comprehension question responses recorded: to better investigate some of the partially supported hypotheses in Phase 1. Here, we'd expect that a human video upper-baseline would receive even higher comprehension scores than animation upper-baseline produced by an animator for the following reasons: First, animating facial expressions accurately is too difficult with the use of current ASL animation technology [Elliott et al. 2008; Filhol et al. 2010; Fotinea et al. 2008; Vcom3D 2012]. Handling complex aspects of facial expressions such as the exact face, timing of the intensity with the hands, simultaneous performance, and transitions is beyond the state of the art of current ASL systems [Huenerfauth et al. 2011]. Thus, what an animator can achieve as an upper baseline might lack the naturalness and the quality of video of a human signer. Second, facial expressions introduce new challenges in achieving that delicate "balancing act" between a natural, fluent version of the stories from the human signer and the control of important variables between the upper-baseline and the animation being evaluated. Finally, deaf viewers tend to focus on signers' faces, as shown in [Emmorey et al. 2008], which could result to an audience sensitive to facial expressions errors.

5.1 Design of Evaluation Studies for Phase 2

In Phase 2 we evaluate a model for performing facial expressions in ASL animation, whose movements of the face depend on the grammar or emotions of the stories being displayed. Facial expressions are a required part of ASL and can differentiate the meaning of identical sequences of hand movements [Neidle et al. 2000]. For this study, we investigated six categories of meaningful facial expressions: yes/no question; rhetorical question; negation; topic; wh-word question (e.g., who, what, why, and how); and emotions (e.g., frustration, sadness and irony). For this article, because we were primarily interested in the affect of different upper baselines, it was not important for the "model" being evaluated to be very sophisticated. Thus, our model was a simplistic rule: apply a facial expression from the above six categories over a whole sentence with the same grammar/meaning, e.g. a y/n-question face over the entire ASL sentence asking a question that can be answered with a yes or no.

We want to evaluate the understandability of ASL animations of three types: (1) a lower baseline consisting of stories with a static face (no facial expressions), (2) animations with facial expressions that follow the simplistic "model" above, and (3) an upper baseline. To evaluate hypotheses H1-H4, we conducted a pair of studies: (i) one in which the upper baseline was a video of a human ASL signer and (ii) one in which the upper baseline was an animation whose facial expressions were produced by a native ASL signer using some animation software [Vcom3D 2012]. Since the same model is being evaluated in both studies, the simplicity of the model does not affect the comparison of the different upper baselines used in each study. The design and conduct of these studies was similar to the pair of studies in Phase 1.

Our studies consisted of two parts: In part 1, participants viewed animations of a virtual human character or human videos telling a short story in ASL. Each story included one of the above six categories of facial expressions (whereas the stories in Phase 1 focused on complex ASL verb signs). Fig. 7 shows a story transcript; the bars with the abbreviations over the script indicate the required facial expression to be performed during some of the signs. An English translation of this story would be: "Alex usually takes Math classes. This semester, the school doesn't have any science classes. Alex is taking two classes." Participants watched each story, which was one of three types: lower baseline (no facial expression), animation with facial expressions based on our simple-rule model, and upper baseline (which was either a human video or an animator-produced animation). Then, participants answered 4 yes/no comprehension questions (Fig. 7). Stories and questions were engineered in such a way that the wrong answers would indicate that the users misunderstood the facial expression displayed. In Fig. 7, if the "negation" facial expression was not noticed by a participant, then the participant would think that the school does not offer science classes, which would affect the participant's answers to the questions. For each story viewed, participants also responded to 1-to-10 Likert-scale questions about how grammatically correct, easy to understand, naturally moving the hands and the face of the animation/human signer appeared. These Likert-scale questions were identical to those used in Phase 1.

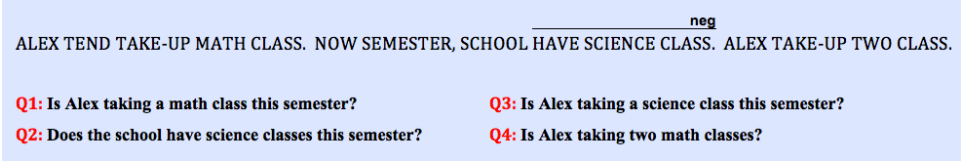

Fig. 7. Example script and corresponding comprehension questions for a story shown in the study.

In part 2 of the studies, participants viewed three versions of a single ASL sentence side-by-side on one screen. The sentences shown side-by-side were identical, except that they were of different versions: lower baseline animation, model-based animation, and upper baseline (either a video or an animation, depending on the study). Fig. 8(a) contains a screenshot of what a participant would see side-by-side in the study with the animation upper baseline, and Fig. 8(b) depicts what was seen in the study with the video upper baseline. The participants could re-play each animation as many times as they wished. Participants were asked to focus on the facial expressions and respond to a 1-to-10 Likert-scale question about the quality of each of the three versions of the sentence. The methodology used here is similar to studies in Phase 1.

For both studies, we engineered a total of 28 ASL stories, distributed as follows: y/n-question (4), rh-question (4), negation (4), topic (4), wh-question (4), and emotions (8). The stories were split in the two parts of the studies in a 3:1 ratio, resulting at 21 stories for the part 1 and 7 stories for the part 2. Beside the upper-baselines used, all the other animations and their sequencing in our pair of studies were identical. A fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 7 animations of each version: i) lower baseline, ii) model, or iii) upper baseline. Again, all of the instructions and interactions for both groups were conducted in ASL by a deaf native signer, who is a professional interpreter. Part of the introduction, included in the beginning of the experiment, and the comprehension questions in part 1 of both studies were presented by a video recording of the interpreter.

Fig. 8. Screenshots of the side-by-side comparison portion of the studies as shown to participants in (a) animator-upper-baseline study and (b) video-upper-baseline study.

5.2 Creation of Human Video Upper Baseline for Phase 2

To produce the human video upper baseline, as done in Phase 1, we recorded a native signer (the same person as in Phase 1) in our studio, sitting on a stool in front of a blue curtain, wearing a green t-shirt. The camera placement was identical to the recording process in Phase 1, and the same monitor was placed below the camera with the story

scripts (like the example story shown in Fig. 7). As before, the signer had time to memorize and practice each of the scripts prior to the recording session. All of the instructions and interactions for the recording session were conducted in ASL by another native signer (same person as in Phase 1) sitting behind the camcorder. The cropping, placement of the signer in the video frame, video size, resolution, and framerate were identical to the animations in this study and the human video in Phase 1.

For the recorded video to serve as an effective upper baseline for comparison, we again wanted to "control" as many of the variables of the ASL performance as possible. For this study, we primarily care about how the facial expressions differ between the upper baseline and our model animation under evaluation; so, we would prefer for the other aspects of the performance to be identical. While in Phase 1, the animations being evaluated pre-existed the video upper baseline (because we were replicating an earlier study), during Phase 2, we were able to record the video upper baseline prior to producing our animations. Thus, the human signer had fewer constraints on the performance because they did not need to mimic the animation. However, producing such a video of a human was still challenging. Because our stories and comprehension questions were carefully engineered, the signer had to perform a specific "script" of signs and the correct category of facial expression during a story, since a difference in the facial expressions could result in a different meaning (e.g. negating a statement). Since ASL has no standard written form, we had to explain our notation scheme (Fig. 7) to the participant being recorded. The stories were somewhat complicated and were engineered to cause confusion if the wrong facial expressions were applied [Kacorri et al. 2013]. They were an average of 9 signs in length, with 1-4 main characters set up at various locations in the signing space, with at least one facial expression per story. So, the human signer required several minutes of practice in order to perform each story smoothly. During the recording process, the signer glanced at the script occasionally; so, the videos include some moments when his eyes glance between the monitor displaying the story transcripts and the camcorder.

While the signer had greater freedom in performance than in Phase 1, we still had to let the signer know about some restrictions, in regard to: (i) pausing between stories and positioning the hands down at a default pose at the beginning and end of each story, (ii) controlling the intensity of the facial expression in the story (so as not to be comically exaggerated in an unnatural manner), and (iii) avoiding embellishments, e.g., additional emotional facial expressions on top of grammatical facial expressions. Since we were not investigating co-occurring facial expressions in our experiment, it would be undesirable for the human signer to add such embellishments to the video. We also asked the signer to avoid using ASL signs that were influenced by English, such as alphabet-letter-initialized signs. After practicing for a few minutes, the signer attempted to perform each story multiple times until we produced an acceptable recording. An average of 3 attempts per story were recorded. From an overset of 39 ASL story scripts used during the recording session, only the 28 ASL stories that we recorded that best met the above criteria were selected for inclusion in the study.

Even though we started with the human signing the scripts, the coaching and scripting process, described in Section 4.2 as a delicate "balancing act," was still challenging for this study. The human signer needed to exercise control over micro-movements of his face, which is an acting skill that is beyond that needed in spontaneous signing. Further, it was difficult for the research team to evaluate the quality of these micro-movements in real time; so, it was necessary to replay videos during the recording process to assess the quality. Moreover, our standards for the video quality were higher in

Phase 2 because we expected our study participants to be very sensitive to unnatural facial expressions. Prior research on deaf native ASL signers has indicated that they visually fixate primarily on the face of the person who is signing [Emmorey et al. 2008].

5.3 Data Collection and Results of Phase 2

Similar methods were used as in Phase 1 to ensure that participants were native ASL signers and that the study environment was ASL-focused with little English influence. Ads were posted on New York City Deaf community websites asking potential participants if they had grown up using ASL at home or had attended an ASL-based school as a young child. The study with the video upper baseline included 18 participants: 15 participants learned ASL prior to age 5, and 8 participants attended residential schools using ASL since early childhood. The remaining 10 participants had been using ASL for over 12 years, 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 9 men and 9 women of ages 22-45 (average age 31.6). The study with the animation upper baseline included 16 participants: 10 participants learned ASL prior to age 5, and 6 participants attended residential schools using ASL since early childhood. The remaining 10 participants had been using ASL for over 9 years, 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 11 men and 5 women of ages 20-41 (average age 31.2).

Fig. 9, 10, and 11 display the results from the video-upper-baseline and animation-upper-baseline studies, including the following response data: comprehension-question scores and Likert-scale scores collected in part 1 of the studies (after a participant viewed a story) and the Likert-scale scores collected in part 2 of the studies (in which participants assigned a score to each of the three sentences viewed side-by-side). Labels ending with the letter "A" indicate data collected in animation-upper-baseline study, and labels ending in the letter "V" indicate data collected in video-upper-baseline study. See Section 4.3 for additional details (not repeated here) about error bars, the pooling together of the Lower and Model data to produce the "Model+Lower" category, layout of the graphs, statistical tests performed, planned comparisons, and the use of stars to indicate statistical significance in the graphs.

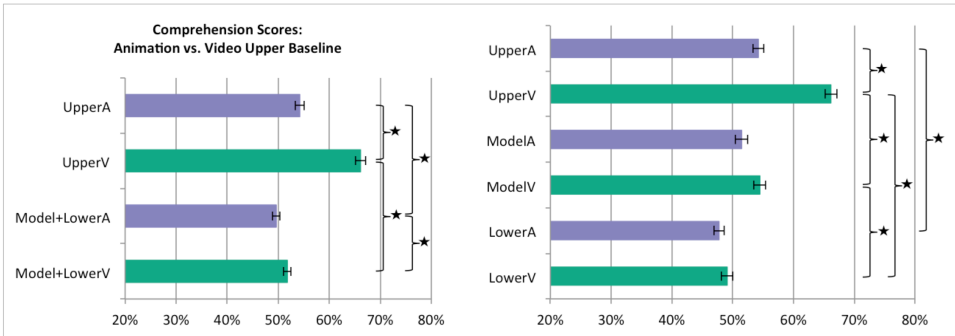

Fig. 9. Results of comprehension scores in Phase 2.

Fig. 9 displays the comprehension-question accuracy scores from part 1 of the studies. We see that there was a significant difference between UpperA and UpperV; so, here hypothesis H1 was supported. This is a different outcome than was observed in Phase 1; here, the videos of a human signer achieved higher comprehension scores than

the animations of a virtual human with the facial expressions carefully animated by a human. Given that handling complex aspects of facial expressions is beyond the state of the art of current ASL synthesis systems, it was not surprising that the upper baseline created by a native ASL animator received lower scores than the human video.

In Phase 2, Hypothesis H2 was not supported. In fact, contrary to our hypothesis, Model+LowerV actually had *significantly higher* comprehension scores than Model+LowerA (Mann Whitney U-test, alpha=0.05). However, the magnitude of this difference was quite modest (approximately 2% higher), and TOST equivalence testing indicated that the values are actually "equivalent" according to our (-15%, +15%) margin. While no story was displayed more than one time during the study, we speculate that seeing a video of a human performing some of the ASL stories may have helped participants grasp the overall genre of the stories in the study. Perhaps participants were able to realize that all of the stories contained a subtle ambiguity that depended on the facial expression and that the facial expressions were of different types (yn-question, emotion, etc.). This may be why there were slightly higher comprehension scores for Model+Lower in the video-upper-baseline study.

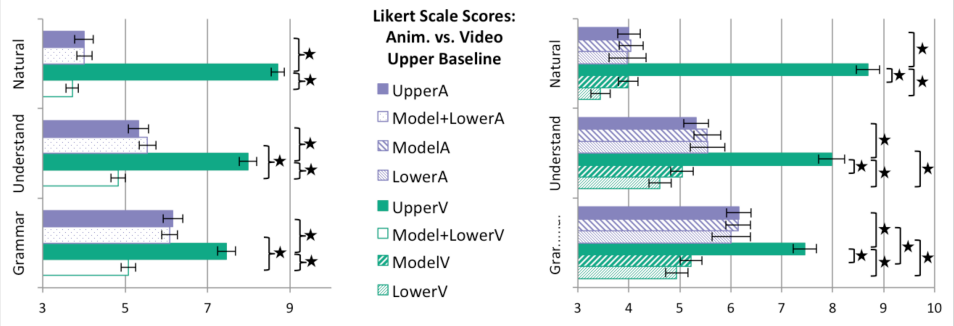

Fig. 10. Results of grammaticality, understandability, and naturalness Likert-Scale scores in Phase 2.

Fig. 10 displays the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness from part 1 of the studies. UpperV had significantly higher grammaticality, understandability, and naturalness scores than UpperA – thereby supporting hypothesis H3, that video upper baseline would get higher subjective Likert-scale scores than an animation upper baseline.

The results in Fig. 10 partially support hypothesis H4, that the use of a video upper baseline would affect the Likert-scale subjective scores for the other stimuli in the study. For grammaticality and understandability, we see a significant difference between "Model+LowerA" and "Model+LowerV." This is a different outcome than was observed in Phase 1, when no significant difference was observed for Likert-scale data in part 1. We speculate that the videos of a human with facial expression was perceived as so much better than animations by our participants in Phase 2 that it may have affected the participants "calibration" of their Likert-scale responses, resulting in lower Likert-scale subjective scores (for grammaticality and understandability) for the non-video stimuli.

Fig. 11 displays the Likert-scale subjective scores collected from participants during part 2 of the studies (the side-by-side comparison of identical sentences, with different versions of the facial expressions that could be replayed multiple times). UpperV is significantly higher than UpperA, further supporting hypothesis H3 that human upper baselines would get higher Likert-scale subjective scores than animation upper baselines.

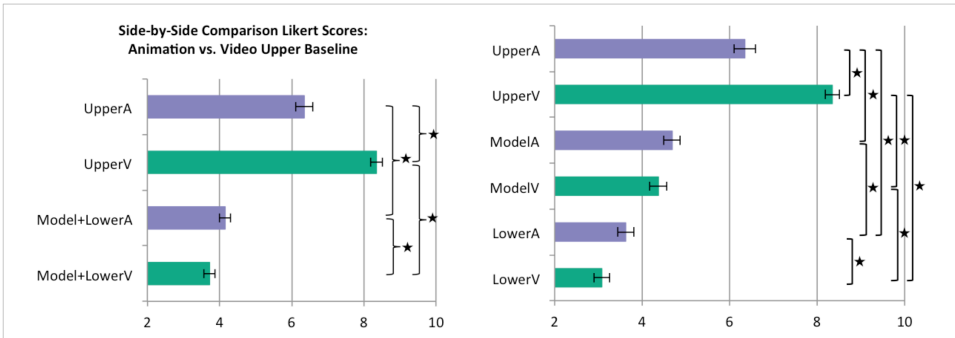

Fig. 11. Results of Side-by-side comparison scores in Phase 2.

In Fig. 11, hypothesis H4 was again supported: Using a human video upper baseline depressed the subjective Likert-scale scores that participants gave to the animations. Mode1+LowerV was significantly lower than Model+LowerA (same for the detailed pairs ModelV/ModelA and LowerV/LowerA). The magnitude of this depression is 10%-20%. As in Phase 1, this is not a surprising result; when looking at videos of humans in direct comparison to animations of a virtual human character, it is reasonable that participants would feel that the animations are less natural/grammatical.

It is notable that we did not observe a significant depression for naturalness in Fig. 10. As mentioned in Section 4.3, we speculate that the depressive effect of displaying video upper baselines may depend on whether participants are assigning Likert-scale subjective scores to videos in a side-by-side direct comparison (as in part 2, Fig. 11) or sequentially throughout a study (as in part 1, Fig. 10). Perhaps in the side-by-side setting, a greater "comparativeness" is triggered in the participants, and the visual distinctness of the video "stands out" in comparison to animations – thereby resulting in a stronger depressive effect on the Likert-scale scores for the other stimuli in the study.

6. PHASE 3: ANIMATION VS. VIDEO COMPREHENSION QUESTIONS

Aside from being used as an upper baseline, a video of a human signer could appear within the software interface presenting animations in a user study (as mentioned at the end of Section 2). Specifically, comprehension questions that participants are asked after viewing each of the sign language animations might be displayed in different modalities (human video or animation). In this study we investigate whether this choice affects the comprehension (Hypothesis H5) and subjective Likert-scale (Hypothesis H6) scores collected in a study with deaf participants. To evaluate these hypotheses, we conducted a final pair of studies: one with comprehension questions presented as high-quality ASL animations and one with comprehension questions presented as human videos.

6.1 Design of Evaluation Studies for Phase 3

We had previously conducted a user study that presented the comprehension questions in the form of animations of a virtual human character [Huenerfauth and Lu 2010]; for that study, a native ASL signer with animation experience carefully produced the animations using sign language animation software [Vcom3D 2012]. For Phase 3, we decided to replicate that study; we replaced the animations containing the comprehension questions with videos of human signer performing identical ASL questions. Because we wanted to isolate the effect of showing a human video for the comprehension questions, we decided to use an animation as the upper baseline in both the 2010 and 2013 study. So, the only difference between the 2010 and 2013 studies is whether the comprehension questions

were presented in the form of: (1) video in which a human signer asked questions in ASL about information contained within the stories being presented or (2) animation in which an virtual human character asked identical questions. It is important to note that the stimuli shown in the two studies were identical; the only difference was how the comprehension questions were presented. Fig. 12(a) illustrates what the participants saw in our animation-comprehension-questions study in 2010 (with an ASL story shown on one slide and four questions displayed on the next slide), and Fig. 12(b) illustrates what participants saw in our video-comprehension-questions study in 2013. We are interested in how the participants' scores on comprehension questions might change (Hypothesis H5), and we are also interested in whether there might be an effect on the scores for the Likert-scale questions about the naturalness, understandability, and grammaticality of the ASL stories (Hypothesis H6). Unlike the studies described in Phase 1 and 2 of this article, we did not conduct the part 2 side-by-side comparison of the stories in our Phase 3 experiment because no aspect of that part of the study differed between 2010 and 2013, since there were no comprehension questions asked in that part of the study. Similar to the study in 2010, participants were asked to view a total of 9 short stories in ASL. Again, a fully-factorial design was used such that: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 3 animations of each version: i) lower baseline, ii) model, or iii) upper baseline.

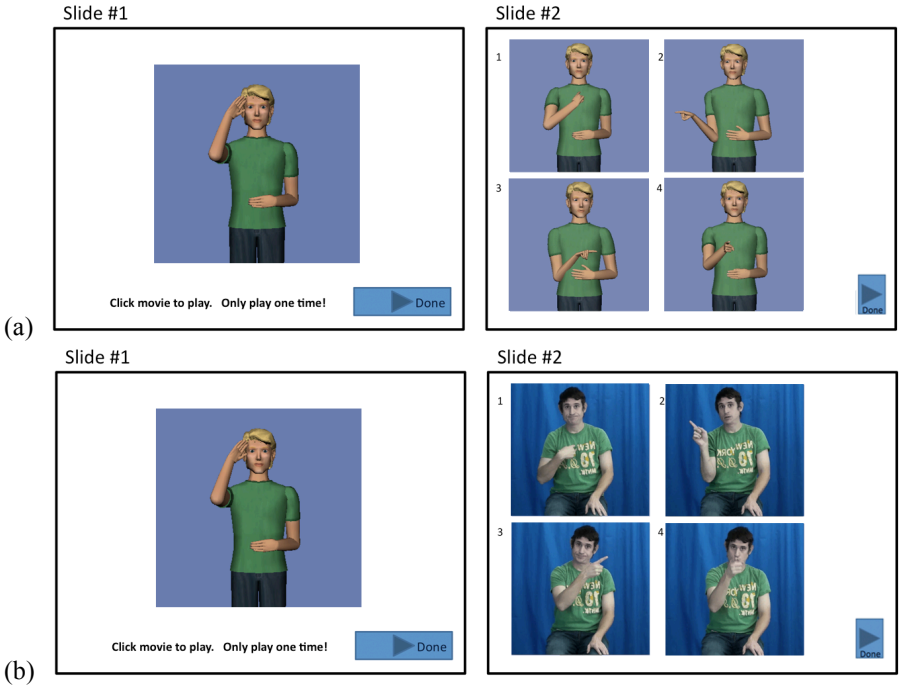

Fig.12 Screenshots of two forms of comprehension questions presented in 2010 study (a) and 2013 study (b).

To produce the human video comprehension questions, we recorded the videos from a native signer, with the same blue background, camera angle, and other details, as described in Section 4.2. We used one large monitor in front of the signer to display the scripts of the comprehension questions (the studio setup is similar as shown in Fig. 3).

6.2 Results of Phase 3

This section presents the results of our 2013 replication of our original 2010 study, and it compares the results of these two studies. The data collected include the comprehension-question and Likert-scale scores after a participant viewed a story one time. Details of the participants in the 2010 study were described in Section 4.3. The 2013 study included 18 participants: 17 participants attended residential schools using ASL since early childhood, and the 18th participant used ASL since birth, attended mainstream schools, and attended a university with instruction in ASL. 15 participants learned ASL prior to age 5. There were 12 men and 6 women of ages 20-37 (average age 28.8).

In the results illustrated in Fig. 13 and 14, the thin error bars display the standard error of the mean. Animator10 and Animator13 were upper baselines, Lower10 and Lower13 were lower baselines, Model10 and Mode113 were versions of the animations produced using our verb inflection model (details in Section 4), and Model+Lower10 and Model+Lower13 were the combined data from the Model and Lower groups. It is important to note that all of the stimuli were identical in 2010 and 2013 (i.e., Animator10 and Animator13, etc.); the only difference in those two studies was that comprehension questions used to collect the evaluation scores were presented in different forms: as animations in the 2010 study and as videos of a human signer in the 2013 study.

As mentioned in section 4.3, since H5 and H6 hypothesize no differences, TOST equivalence testing was performed. With alpha=0.05, 90% confidence intervals were calculated (via t-tests for comprehension question scores and via Mann-Whitney U-tests for Likert-scale scores) for the following pairs of values: (a) Animator13 and Animator10, (b) Model+Lower13 and Model+Lower10, (c) Model13 and Model10, and (d) Lower13 and Lower10. Section 4.3 explained how we selected an equivalence margin interval of (-1.5,+1.5) for Likert-scale scores and (-0.15,+0.15) for comprehension question scores. According to the TOST procedure, whenever the entire confidence interval falls within our equivalence margin interval, then the pair of scores is deemed equivalent.

While it was not necessary for examining H5 and H6, we also conducted one-way ANOVAs (for comprehension question scores) and Mann-Whitney U-tests (for Likert-scale scores) for the following pairs of values: (1) all three values from 2010, (2) all three values from 2013, (3) Animator13 and Model+Lower13, and (4) Animator10 and Model+Lower10. Statistical significance ($p<0.05$) for any of these comparisons has been marked with a star in Fig. 13 and 14. As discussed in section 4.3, while not necessary for testing our hypotheses, we believe it is useful for us to present these statistical tests in our paper, for reference, since future researchers evaluating the quality of an animation relative to upper and lower baselines would traditionally perform these comparisons.

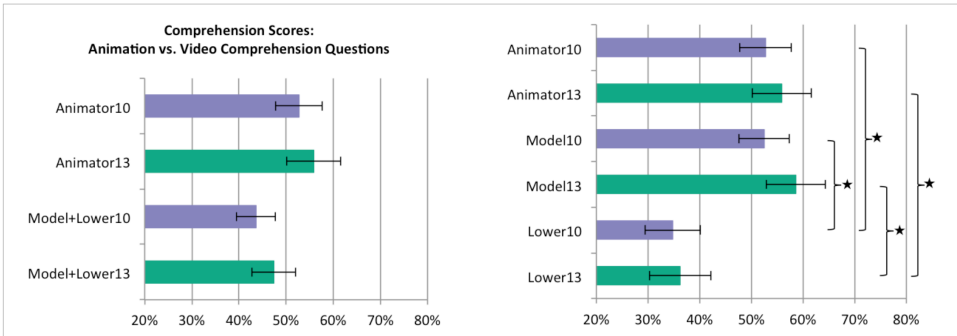

Fig. 13. Results of comprehension scores in Phase 3.

Hypothesis H5 would predict that the mode of presentation of the comprehension questions (animation vs. video) would not affect the comprehension scores collected in the study. This was mostly supported by the results. TOST equivalence testing indicated that the following pairs of values were equivalent: Animator10 and Animator13, Lower10 and Lower13, and Model+Lower10 and Model+Lower13. (The results were inconclusive for Model10 vs. Model13.) It should be noted that the animations used in 2010 to present the comprehension questions were carefully produced by a human animator and were of good quality; we predict that if low-quality animations had been used, then we would have seen lower comprehension scores in 2010 (due to confusion from participants who did not understand the question being asked).

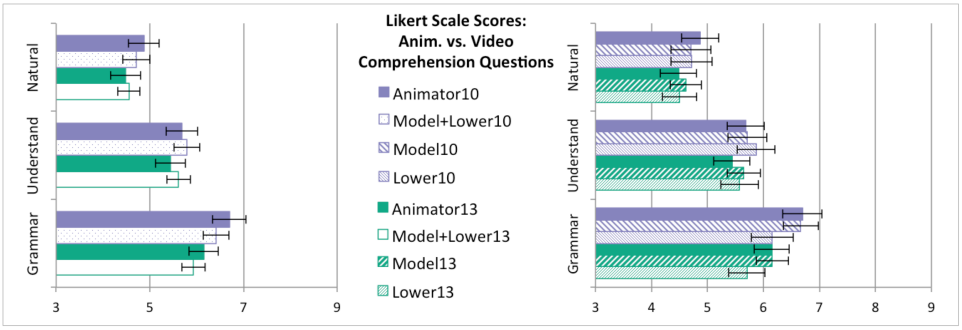

Fig.14 Results of grammaticality, understandability, and naturalness Likert-scale scores in Phase 3.

Fig. 14 illustrates the 1-to-10 Likert-scale subjective scores for grammaticality, understandability, and naturalness from the studies in Phase 3. Hypothesis H6 would predict that the mode of presentation of the comprehension questions (video vs. animation) would have no effect on the Likert-scale subjective evaluation scores in the study, and this was mostly supported by TOST equivalence testing. (The scores for Naturalness for Animator10 and Animator13 were inconclusive; all other compared values were determined to be equivalent.) Thus, we conclude that H6 was supported.

7. CONCLUSIONS AND FUTURE WORK

This article has investigated several methodological issues that are important for researchers in the growing field of sign language animation research, who are conducting experimental studies with sign language users evaluating their animations. Specifically, we examined whether certain choices in experiment design affect the comprehension and subjective scores collected in a study. We quantified the effects of changing the mode of presentation (video vs. animation) of two elements of a study: the upper baseline stimuli and the comprehension questions. Awareness of such effects is important so that future researchers can make informed choices when designing new studies and so that they can fairly compare their results to previously published studies, which may have made different methodological choices.

In order to investigate these issues, we conducted several replications of experiments in which most of the stimuli were held constant, and we were able to measure if there was a difference in the scores collected from participants when we changed the upper baseline or modality of presentation for the comprehension questions. To make our results useful to a wide variety of researchers, we included a variety of study designs and question formats, including: comprehension questions responses, Likert-scale subjective scores of a single stimulus, and Likert-scale subjective scores of multiple stimuli presented side-by-side. The three pairs of experiments conducted allowed us to evaluate six hypotheses:

- H1: Video upper baselines received higher comprehension scores than animation upper baselines. This difference was significant in Phase 2 experiments, which included ASL animations with facial expressions, but it was not significant in Phase 1. We speculate that the effect may occur when there is a greater quality-difference between the animations and videos, as there is when they include facial expressions, which are not handled well by current animation tools. An alternative may be that the effect could be observed in Phase 2 because the studies contained more stories, and thus a larger number of comprehension question responses were collected.
- H2: When video upper baselines were used instead of animation upper baselines, the comprehension question scores for the other stimuli in the study increased slightly. This increase was significant in Phase 2, but it was inconclusive in Phase 1. We speculate that the video upper baseline shown during the study may have given the participants some additional advantage in answering the comprehension questions for the other stimuli because it allowed them to better understand the genre of stories being presented, the relationship between the stories and the comprehension questions, or the types of facial expressions that they should expect to see in the other (animation) stimuli. We had not hypothesized that we would observe such an effect: H2 had been that changing the upper baseline from animation to video would have *no* effect on the comprehension question scores for the other stimuli. The presence of this effect may depend on the degree to which participants in the study are able to "learn something useful" from watching the high-quality video upper baseline stimuli that generalizes to the other stimuli in the study; thus it may be magnifying a "learning effect" that was inherent to the design of a study.
- H3: Video upper baselines received higher Likert-scale subjective scores than an animation upper baseline. This hypothesis was supported by statistically significant differences observed in both Phase 1 and Phase 2. This effect was present in both Likert-scale subjective scores collected after sequential presentation of an individual stimulus (as done in part 1 of the studies in Phases 1 and 2) and in Likert-scale subjective scores collected during simultaneous presentation of multiple stimuli side-by-side (as in part 2 of the studies). Given the state of the art of sign language animation technologies (and that there are still many unsolved challenges to address in the field), it is not surprising that videos of humans would receive higher subjective evaluation ratings than animations.
- H4: This hypothesis is best considered if it is split into two sub-cases: (H4a) for sequential presentation of stimuli, as in part 1 of the studies in Phase 1 and 2, and (H4b) for simultaneous side-by-side presentation of stimuli, as in part 2.
  - H4a: When video upper baselines were used instead of animation upper baselines, the Likert-scale subjective evaluation scores for the other stimuli in the study decreased during sequential presentation. This difference was significant for some of the Likert-scale categories in Phase 2 (grammaticality and understandability), but it was not significant for any of the categories in Phase 1. So, sub-hypothesis H4a is partially supported by the results presented in this article – during the experiments with sign language animations with facial expressions.
  - H4b: When video upper baselines were used instead of animation upper baselines, the Likert-scale subjective evaluation scores for the other stimuli in the study decreased during simultaneous (side-by-side) presentation of stimuli. This difference was significant in both Phase 1 and Phase 2. As discussed earlier, we speculate that participants felt a greater sense of "comparativeness" when the

stimuli were shown side-by-side, and this may have strengthened the depressive effect on the Likert-scale scores for the animation stimuli when a video upper baseline was shown.

- H5: Displaying comprehension questions as videos of a human signer or as a high-quality animation will not affect the comprehension questions accuracy scores. Statistically equivalent comprehension scores were collected in the studies with video or with high-quality animations used to present comprehension questions in Phase 3.
- H6: Displaying comprehension questions as videos of a human signer or as high-quality animations will not affect the subjective Likert-scale scores that participants assign to the animations in the study. The choice of video or high-quality animation led to statistically equivalent Likert-scale scores collected in both studies in Phase 3.

The two main contributions of this article are: (a) providing methodological guidance for future researchers who are conducting studies with sign language and (b) facilitating fair comparisons of the results of sign language animation evaluation studies, in which the authors have made different methodological choices.

7.1 Recommendations for Future Researchers

This section discusses how the conclusions outlined above can be translated into concrete methodological guidance for researchers conducting evaluation studies with sign language animations. First of all, while not a major focus of this article, the conclusions above and our prior research [Huenerfauth et al. 2008] have indicated that comprehension question scores and Likert-scale subjective evaluation scores for sign language animations often do not have identical results, and we recommend to future researchers that they include both forms of evaluation in any future studies, since they may be measuring different aspects of sign language animations. Of course, the primary focus of this article has been on the mode of presentation for two aspects of a sign language evaluation study: the upper baseline and the comprehension questions. In particular, two forms of presentation were examined: videos of human signers and high-quality animations of a virtual human. This article does not give a single "correct answer" for the best choice that future researchers should make when designing their studies; indeed, either choice (video or animation) is potentially valid. The selection should be based on the research goals of the study, the practical challenges in producing animations or videos, and the expected effect of these methodological choices on the data collected.

Goals of the Study

Researchers considering which form of upper baseline to use should consider the research questions they want to explore. Given our slightly different results in Phase 1 and Phase 2 of this article, researchers may want to consider whether animations in their study are closer to those in Phase 1 (hand movements during verb signs) or Phase 2 (facial expressions), when considering our results. Each choice of upper baseline has trade-offs that must be considered in regard to the requirements of the study design:

- Video upper baselines would be preferable for researchers studying computer graphics issues relating to the visual appearance of a virtual human for sign language animations, since this would serve as an "ideal" of photorealism.
- Researchers who want to convey to a lay audience the overall understandability of their sign language animations (i.e., the current state of the art) may wish to use videos of humans as an upper baseline (because they are more familiar than animations as a basis for comparison). Of course, researchers would need to

explain the limitations of current sign language animation technologies to manage the expectations of a lay audience being presented their results.

- Researchers who are studying particular linguistic aspects of sign language animations (e.g., the speed/timing of signs or the timing of facial micro-expressions in relation to hand movements) may find an animated-character baseline more useful to their research because it is possible to control the variables of the character's movements precisely. A human in a video may not be able to voluntarily and consistently control these aspects of the performance as necessary for a study.
- For study designs in which it is important that the participants cannot easily determine which stimuli are the upper baselines, animation (with a virtual human identical in appearance to the one in the synthesized animations) are desirable.

Challenges in Producing Videos or Animation

Researchers should also consider the practical challenges they may face in producing videos or animations for use as upper baselines or as comprehension questions:

- *Challenges in producing video upper baseline:* We found that it is harder than many researchers may expect to produce a human video that is a good upper baseline (see Sections 4.2 and 5.2). Scripting and coaching was needed to ensure that our human videos had the same sign sequence, point locations, facial expressions, speed, and other performance variables as our other stimuli. If the study requires that some performance variable is held constant that is very detailed (e.g. precise millisecond timing of speed/pauses, exact height of the eyebrows, etc.), then this may be too difficult for a human to perform voluntarily and consistently. To avoid producing an artificial-looking result, a delicate "balancing act" (as discussed in Section 4.2) was needed between controlling the human's performance and providing freedom so that the result is fluent and natural. The researchers must decide what level of embellishments or improvisation they will tolerate from the human signer. If they prevent the signer from performing aspects of signing (because those aspects are outside the repertoire of the animation system being evaluated), then the video upper baselines may be artificially limited in how natural/fluent they appear.
- *Challenges in producing animation upper baseline:* There are also challenges in producing a good-quality animation that is an effective upper baseline. Depending on the specific animation system/tool used by the researchers, the ease with which a human animator can control their virtual human to produce high-quality signing may vary. If it is possible to blend software-controlled with human-animator-controlled elements of the performance for the virtual human, then it may be easier to produce an upper baseline with variables that are held constant between the upper baseline and the other stimuli. In our studies, we found that our animation tool made it easy for the human animator to add novel hand movements, but the set of facial expression controls was limited. This may have resulted in some of our upper baseline animations in Phase 2 having lower quality than we would have preferred.
- *Challenges in producing video comprehension questions:* While we did not find it especially difficult to produce video recordings of a human signer performing comprehension questions for use in our study in Phase 3, we did need to provide

scripts and coaching during the process. Since there is no standard written form for ASL, it was necessary to explain our script notation to the signer. Further, there are regional/dialectical variations in how certain signs are performed in ASL, and we needed to ensure that the same variant of a sign used in our ASL story stimuli was used during the comprehension questions, to avoid confusion during the study.

- *Challenges in producing animation comprehension questions:* In Phase 3, there was no significant difference in the comprehension or Likert-scale scores when we presented our comprehension questions either as videos or as animations. However, our comprehension question animations were high-quality animations produced by a native ASL signer who had experience using our animation tool [Vcom3D 2012]. If future researchers were using an animation tool of lower quality (or asking comprehension questions that required some linguistic/performance aspect that was beyond the repertoire of their animation tool), then the resulting comprehension questions may be difficult for participants to understand. In that case, we would expect that the comprehension question scores collected in the study would be lower, due to the participants' difficulty in understanding the question being asked.

Effects on Collected Scores from Methodological Choices

Future researchers may also consider how the responses they collect will be affected by their choice of upper baselines and the mode of presentation for comprehension questions. Given that video upper baselines received higher comprehension and Likert-scale scores than animation upper baselines in our studies (Hypotheses H1 and H3), researchers should expect that if they use a video, their upper baseline scores would be higher. Thus, the other stimuli might appear relatively worse by comparison (to a naïve reader of their study who only considers the relative values of the raw scores). Similarly, when using videos as an upper baseline, we observed a depressive effect on the Likert-scale scores for the other stimuli in the study during side-by-side comparisons and, in some cases, during sequential evaluation of stimuli (Hypothesis H4).

Given that researchers may have an interest in the sign language animations they synthesize appearing more successful, this might suggest that there is an incentive for researchers to avoid video upper baselines. Given the advantages of video upper baselines for some study designs and the ease-of-interpretation of the results (mentioned in the "Goals of the Study" section above), it would be inappropriate to avoid the use of video upper baselines merely because they may make animations appear less understandable by comparison. It is for this reason that we believe methodological studies such as this article are important for the research community because it can provide a resource for future researchers who can explain how the results of their study should be interpreted when video upper baselines have been used for comparison.

While the use of video upper baselines clearly led to larger differences in the Likert-scale scores between the upper baseline and the other stimuli (since the upper baseline scores were higher and the other stimuli were lower), the results were more complex for comprehension question scores. In Phase 2, we observed an across-the-board increase in comprehension question scores for all of the stimuli in the study, when video upper baselines were used (Hypothesis H2). In our discussion of those results, we speculated that the result could have been due our participants learning something from watching the video upper baseline that generalized to the other stimuli in the study. Future researchers

designing studies in which there could be a similar learning effect may see a similar resulting increase in comprehension scores when video upper baselines are used.

Based on the results in Phase 3, we did not see any significant difference in the scores collected in studies that used video or used animation to present comprehension questions. However, as noted above, this was evaluated using animations that were high-quality, and researchers may see lower comprehension question scores if they use difficult-to-understand animations to present their comprehension questions in a study.

7.3 Future work

While this article has focused on ASL and participants in the U.S., the results should also be relevant to researchers studying other sign languages internationally. Further, we believe that the methodology of this article (replicating a user-study with identical stimuli to test the impact of methodological choices) could itself be replicated by researchers studying other sign languages if they wanted to quantify the effects that we observed for ASL, for other sign languages. In future work, we want to provide more guidance for researchers as to the best approach to use when designing their evaluation studies of sign language computer animations. It would be useful to explore the effect of other modes of presentation (e.g., English text, in-person signing), question formats, or tasks that were not evaluated in this article but have been used by prior researchers when conducting studies (Section 2). The maturing field of sign language animation synthesis can benefit from additional methodological research – especially work that facilitates comparisons across evaluation studies of different animation systems or leads to further consensus in evaluation techniques – and this will ultimately benefit the users of this technology.

ACKNOWLEDGEMENTS

Jonathan Lamberton and Miriam Morrow assisted with the recruitment of participants and the conduct of experimental sessions; they also provided several valuable linguistic insights about the performance of ASL inflecting verbs and facial expressions.

REFERENCES

AHLBERG, J., PANDZIC, I.S., YOU, L. 2002. Evaluating face models animated by MPEG-4. In I.S. Pandzic, R. Forchheimer (eds.), *MPEG-4 facial animation: the standard, implementations and applications*, Wiley & Sons, 291–296.

BALDASSARRI, S., CEREZO, E., AND ROYO-SANTAS, F. 2009. Automatic translation system to Spanish Sign Language with a virtual interpreter. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I (INTERACT '09)*, T. Gross, J. Gulliksen, P. Kotz, L. Oestreicher, P. Palanque, R. O. Prates, and M. Winckler (Eds.). Springer-Verlag, Berlin, Heidelberg, 196-199.

BALDASSARRI S., AND CEREZO, E. 2012. Maxine: Embodied conversational agents for multimodal affective communication, *Computer Graphics*, Prof. Nobuhiko Mukai (Ed.), ISBN: 978-953-51-0455-1, InTech.

BERGMANN, K. 2012. The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents. Dissertation, Bielefeld University, Germany.

DAVIDSON, M. J., ALKOBY, K., SEDGWICK, E., BERTHIAUME, A., CARTER, R., CHRISTOPHER, J., CRAFT, B., FURST, J., HINKLE, D., KONIE, B., LANCASTER, G., LUECKING, S., MORRIS, A., MCDONALD, J., TOMURO, N., TORO, J. AND WOLFE, R. 2000. Usability testing of computer animation of fingerspelling for American Sign Language. Presented at *the 2000 DePaul CTI Research Conference*, Chicago, IL, November 4, 2000.

EMMOREY, K., THOMPSON, R., COLVIN, R. 2008. Eye gaze during comprehension of American Sign Language by native and beginning signers. *J. Deaf Stud. Deaf Educ.* (2009) 14 (2): 237-243.

ELLIOTT, R., GLAUERT, J., KENNAWAY, J., MARSHALL, I., SAFAR, E. 2008. Linguistic modeling and language-processing technologies for avatar-based sign language presentation. *Univ Access Inf Soc* 6(4), 375-391. Berlin: Springer.

FILHOL, M., DELORME, M., BRAFFORT, A. 2010. Combining constraint-based models for Sign Language synthesis. In *Proc. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Language Resources and Evaluation Conference (LREC),* Valetta, Malta.

FOTINEA, S.E., EFTHIMIOU, E., CARIDAKIS, G., KARPOUZIS, K. 2008. A knowledge-based sign synthesis architecture. *Univ Access Inf Soc* 6(4):405-418. Berlin: Springer.

GARAU, M., SLATER, M., BEE, S., AND SASSE, M. A. 2001. The impact of eye gaze on communication using humanoid avatars. In *SIGCHI'01*, Seattle, USA. ACM, NY, USA.

GIBET, S., COURTY, N., DUARTE, K., AND LE NAOUR, T. 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 pgs.

HAM, R.T., THEUNE, M., HEUVELMAN, A. AND VERLEUR, R. 2005. Judging Laura: perceived qualities of a mediated human versus an embodied agent. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist (Eds.), (Vol. 3661, pp. 381-393). Springer Berlin / Heidelberg.

HUENERFAUTH, M. 2006. Generating American Sign Language classifier predicates for English-to-ASL machine translation. Doctoral dissertation, Computer and Information Science, University of Pennsylvania.

HUENERFAUTH, M., HANSON, V. 2009. Sign language in the interface: access for deaf signers. In C. Stephanidis (ed.), *Universal Access Handbook*. NJ: Erlbaum. 38.1-38.18.

HUENERFAUTH, M., LU, P. 2012. Effect of spatial reference and verb inflection on the usability of American sign language animation. In *Univ Access Inf Soc*. Berlin: Springer.

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):1-27.

HUENERFAUTH, M., LU, P. 2010. Modeling and synthesizing spatially inflected verbs for American Sign Language animations. In *Proceedings of The 12th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2010),* Orlando, Florida, USA. New York: ACM Press.

HUENERFAUTH, M., LU, P., AND ROSENBERG, A. 2011. Evaluating importance of facial expression in American Sign Language and Pidgin Signed English animations. In *Proceedings of The 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2011),* Dundee, Scotland. New York: ACM Press.

KACORRI, H., LU, P., HUENERFAUTH, M. 2013. Evaluating Facial Expressions in American Sign Language Animations for Accessible Online Information. In C. Stephanidis and M. Antona (eds.), *UAHCI/HCII 2013, Part I, Lecture Notes in Computer Science 8009*, Heidelberg: Springer, 510-519.

KENNAWAY, J. R., GLAUERT, J. R. W. AND ZWITSERLOOD, I. 2007. Providing signed content on the Internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact*. 14, 3, Article 15 (September 2007).

KIPP, M., HELOIR, A., NGUYEN, Q. 2011a. Sign language avatars: animation and comprehensibility. In H. Vilhjálmsson, S. Kopp, S. Marsella, K. Thórisson (eds.), *Intelligent Virtual Agents* (Vol. 6895). Springer, 113-126.

KIPP, M., NGUYEN, Q., HELOIR, A., AND MATTHES, S. 2011b. Assessing the deaf user perspective on sign language avatars. In *Proceedings of ASSETS'11*, Dundee, Scotland. ACM, New York, NY, USA, 107-114.

LÓPEZ-COLINO, F., COLÁS, J. 2011. The Synthesis of LSE classifiers: from representation to evaluation. j-jucs 17(3), 399-425.

LU, P., HUENERFAUTH, M. 2011. Synthesizing American Sign Language spatially inflected verbs from motion-capture data. *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT),* in conjunction with *ASSETS 2011*, Dundee, Scotland.

LU, P., HUENERFAUTH, M. 2010. Collecting a motion-capture corpus of American Sign Language for data-driven generation research. *Proceedings of the First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010),* Los Angeles, CA, USA.

LU, P., HUENERFAUTH, M. 2012. Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation. Manuscript submitted for publication.

LU, P., HUENERFAUTH, M. 2012. Learning a vector-based model of American Sign Language inflecting verbs from motion-capture data. *Proceedings of the Second Workshop on Speech and Language*

Processing for Assistive Technologies (SLPAT), Human Language Technologies: The 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2012), Montreal, Canada.

LU, P., AND KACORRI, H. 2012. Effect of presenting video as a baseline during an American Sign Language Animation User Study. In *Proceedings of The 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2012)*, Boulder, Colorado. New York: ACM Press.

MCDONNELL, R., JÖRG, S., MCHUGH, J., NEWELL, F., AND O'SULLIVAN, C. 2008. Evaluating the emotional content of human motions on real and virtual characters. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization (APGV '08)*. ACM, New York, NY, USA, 67-74.

MITCHELL, R., YOUNG, T., BACHLEDA, B., & KARCHMER, M. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies*, 6(3):306-335.

MORIMOTO, K., KUROKAWA, T., KENTAROU, U., TERUYO, K., KAZUSHI, T., KATSUO, N., TAMOTSU F. 2003. Design of an agent to represent Japanese Sign Language for hearing-impaired people in stomach x-ray inspection. *Asia Design Conference* 2003.

MORIMOTO, K., KUROKAWA, T., AND KAWAMURA, S. 2006. Improvements and evaluations in sign animation used as instructions for stomach x-ray examination. In *Proceedings of the 10th international conference on Computers Helping People with Special Needs (ICCHP'06),* K. Miesenberger, J. Klaus, W. L. Zagler, and A. I. Karshmer (Eds.). Springer-Verlag, Berlin, Heidelberg, 607-614.

NEIDLE, C., D. KEGL, D. MACLAUGHLIN, B. BAHAN, R.G. LEE. 2000. The syntax of ASL: functional categories and hierarchical structure. Cambridge: MIT Press.

OW, S.H. 2009. User evaluation of an electronic Malaysian Sign Language dictionary: e-Sign dictionary. In *Proceedings of Computer and Information Science*, 34-52.

PRAŽÁK, M., MCDONNELL, R. AND O'SULLIVAN, C. 2010. Perceptual evaluation of human animation timewarping. In *ACM SIGGRAPH ASIA 2010 Sketches (SA 2010).* ACM, New York, NY, USA, Article 30, 2 pages.

RUSSELL, M., KAVANAUGH, M., MASTERS, J., HIGGINS, J. AND HOFFMANN, T. 2009. Computer-based signing accommodations: comparing a recorded human with an avatar. *Journal of Applied Testing Technology, 10* (3), 21.

SCHNEPP, J. AND SHIVER, B. 2011. Improving deaf accessibility in remote usability testing. In *Proceeding of ASSETS'11*, Dundee, Scotland. ACM, New York, 255-256.

SCHNEPP, J., WOLFE, R. AND MCDONALD, J. 2010. Synthetic corpora: a synergy of linguistics and computer animation. *Fourth Workshop on the Representation and Processing of Sign Languages, LREC 2010.* Valetta, Malta.

SCHNEPP, J., WOLFE, R., SHIVER, B., MCDONALD, J. AND TORO, J. 2011. SignQUOTE: a remote testing facility for eliciting signed qualitative feedback. *2nd Int'l Workshop on Sign Language Translation & Avatar Technology*, Dundee, UK.

SCHUIRMANN, D.J. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *J Pharmacokin Biopharm*, 15:657–680. doi: 10.1007/BF01068419.

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

VCOM3D. 2012. Homepage. http://www.vcom3d.com/

Received January 2013; revised June 2013; accepted ____