SignQUOTE: A Remote Testing Facility for Eliciting Signed Qualitative Feedback

Jerry Schnepp[1], Rosalee Wolfe[1], Brent Shiver[1], John C. McDonald[1], Jorge Toro[2]
[1]School of Computing, DePaul University, Chicago, IL
[2]Computer Science Department, Worchester Polytechnic Institute, Worcester, MA
{jschnepp, rwolfe, bshiver, jmcdonald}@cdm.depaul.edu, jatoro@wpi.edu

ABSTRACT

Sign synthesis is still an evolving technology and improving it requires the elicitation of qualitative feedback from users. Current options for acquiring qualitative feedback are limited. Face-to-face tests conducted in sign language are expensive. On the other hand, remote tests do not use the preferred language of the test participants. A new tool, SignQUOTE, (Signed Qualitative Usability Online Testing Environment) is a configurable, cross-platform remote testing system based entirely on sign language. It includes an innovative method for capturing qualitative feedback in sign language via webcam. In a comparison study, participants viewed animations of American Sign Language and gave suggestions for improvement. The authors conducted a study using SignQUOTE that presented stimuli identical to those used in a previously-conducted face-to-face study. When comparing the two approaches, the authors found results that are consistent with previous comparison studies of remote and face-to-face testing. SignQUOTE comes with a TestDesigner that allows researchers to customize tests. The software is available as open source.

1. INTRODUCTION

Sign synthesis technology is still in a formative stage, as no system has yet to produce animations whose appearance is preferable to video recordings of a human signer. Improving sign synthesis requires regular feedback from people who use Sign as their preferred language. As noted by Hix and Hartson [1], qualitative feedback is particularly useful in this formative stage, where the goal is to improve the technology. As Ebling and John [2] aptly state, "To be effective, evaluation cannot simply answer with a "yes" or "no" (e.g., "the interface is not usable"), but must provide detailed information about why the design does not work as anticipated or, at least, what problems users experience". This applies to sign synthesis as well.

2. AN IDEAL TEST SETUP

For gathering qualitative data, an ideal test setup is a face-to-face environment [3], where the test is conducted in the preferred language of the participants [4][5]. This includes all questionnaires, the informed consent, and any instructions. If the preferred language of the test population is a sign language, then the test should be conducted in that sign language. Ideally, the test facilitator should be fluent in the same sign language. If hearing note takers are gathering the qualitative feedback, sign language interpreters are necessary to voice the participants' responses. Figure 1 shows the test setup and Figure 2 shows a test in progress.

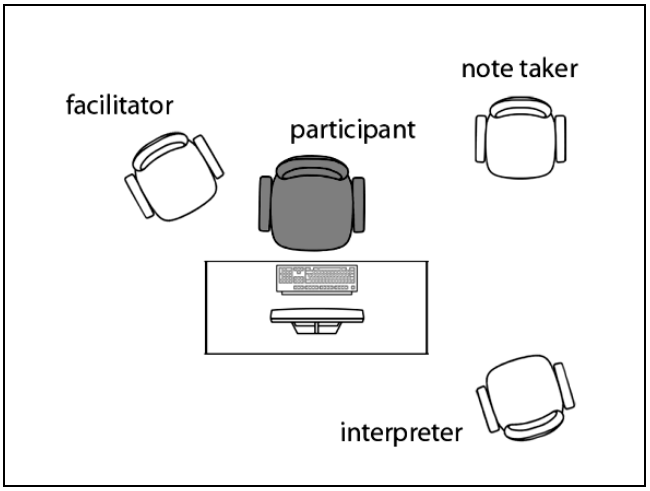

Figure 1: Ideal setup for face-to-face testing.

**Figure 2: Face-to-face testing.
Interpreter is behind camera [6].**

In the United States, the preferred language of the Deaf community is American Sign Language (ASL), an independent natural language that is different from English [7]. For the remainder of this paper, the term "Deaf" (with a capital D) will refer to any community of people whose preferred language is visual/gestural, rather than spoken/written.

SLTAT'11, October 23,2011, Dundee, Scotland.

3. FACE-TO-FACE CHALLENGES

The barriers in face-to-face testing with Deaf participants are myriad. A facilitator not fluent in Sign hampers communication, because an interpreter must repeat everything that the facilitator says. In this situation participants must watch the interpreter, not the facilitator, and any supportive nonverbal cues are lost.

Scheduling is another barrier. In addition to coordinating the schedules of facilitators, note takers and participants, researchers must coordinate with an interpreting agency to schedule certified interpreters. Costs of hiring an interpreter can tend to limit the number of tests.

It can be difficult to recruit enough people that fit the desired user profile and who are willing to incur the time and cost of traveling to the test site. Even in areas where there is a large local Deaf community, the simple challenge of finding and paying for parking can prevent participation. Due to these difficulties, researchers often need to seek out and travel to Deaf conventions.

With these barriers, testing can occur only rarely. In the past, our group's activities were timed to the scheduling of regional Deaf events, in order to attract enough participants.

Even when sufficient data were collected successfully, there was always a potential for problems with locality. Face-to-face testing in a fixed location restricts the pool of potential participants to a specific geographic area. Recruiting exclusively from a local region might skew results when compared with a more geographically diverse population.

4. THE PROMISE OF REMOTE TESTING

In contrast to a face-to-face setting, Web-based remote testing is not limited to a single geographic region. It can be performed asynchronously which eases the burden of scheduling [8], and has been used in recent years to evaluate Web sites, virtual prototypes, and software [9]. This technology can remove barriers of distance, and ease localization problems [10]. Data collected over a network is stored centrally, and testing can occur in parallel, leading to faster data collection and lower costs [11].

Remote testing has the potential to reach a large, geographically diverse Deaf population in a cost-effective manner [12]. It holds particular promise since many members of the Deaf community have embraced the Internet as a preferred means of communication [13]. Through the use of webcams, members of the Deaf community chat directly in Sign and avoid the necessity of typing.

However current remote testing technologies present a significant barrier to eliciting qualitative feedback because they do not permit Deaf participants to respond in their preferred language. In the U.S., English is not a viable option because the average reading fluency of a Deaf adult is at the fourth-grade level [13]. ASL is the preferred language of the Deaf community, and differs radically from English. Asking Deaf participants to type responses to open-ended questions in English forces them to make their suggestions in a second language. This language barrier motivates a new approach to remote usability testing.

5. A DESIGN FOR BETTER ELICITATION

What is needed is an improved approach that would retain the cost savings and convenience of remote testing while providing an easier way for participants to offer qualitative feedback. Desirable features for such a system include

1. *A visual format to the interface.* The only language that should appear is signed language, not written language. It should also minimize the use of graphics, to avoid misinterpretation due to cultural differences.
2. *A facility to record via webcam.* With webcam recording, Deaf participants can answer open-ended questions and thus supply qualitative feedback in their preferred language.
3. *Sequential navigation.* Since no facilitator will be present during the test, navigation should be as simple as possible.

From a researcher's perspective, the system should have minimal hardware requirements and be compatible across platforms, to make it accessible to the widest possible audience. Data should be collected transparently and automatically from each participant and stored in a neutral format. Lastly, creating test designs should require as little technical knowledge as possible.

6. SignQUOTE: A NEW TECHNOLOGY

SignQUOTE (Signed Qualitative Usability Online Testing Environment) is a configurable, cross-platform remote testing system based entirely on signed language. With SignQUOTE, scheduling becomes asynchronous. Participants can test at their convenience by clicking on a URL, and multiple people can participate simultaneously. SignQUOTE makes it possible to invite participants from a wide geographic area, and allows them to test in a familiar setting of their choosing.

It reduces interpreter costs. In a face-to-face setting, interpreters sign the informed consent and test instructions, and need to wait as the participant observes a stimulus and formulates a response. When using SignQUOTE, researchers can wait until all of the tests are completed, and hire an interpreter to voice the open-ended responses in a single session. We found that studies previously requiring sixteen hours of interpreter time now take well under three.

6.1 System Architecture

Figure 3 shows SignQUOTE's two components: TestDesigner and TestServer. A researcher uses TestDesigner to create the test and then directs TestServer to make it available via a hyperlink.

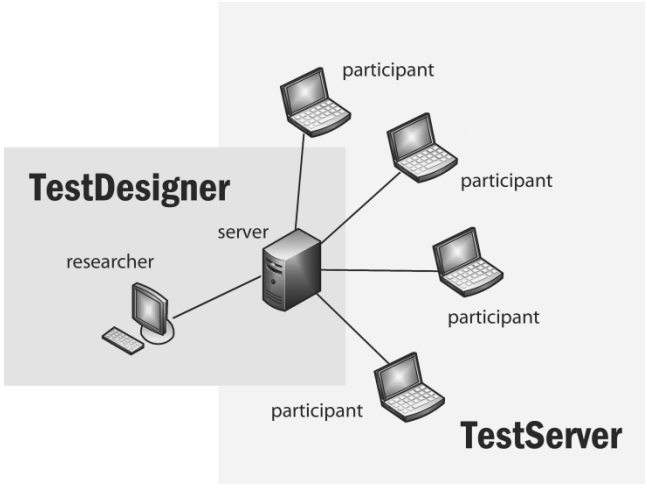

Figure 3: A researcher sets up a test via TestDesigner.

6.2 TestDesigner

The SignQUOTE TestDesigner allows researchers to create, manage, and deploy video-based tests over the Web. It features a graphical user interface with text-based instructions that allows a researcher to easily create and edit questions. Researchers have

the option of recording instructions directly via webcam, or uploading them from pre-recorded video files.

Researchers can specify any number of questions. Choices for question formats include Likert, true/false, or open-ended. Test stimuli as well as instructions can be recorded using a webcam or uploaded as video files. Figure 4 shows a screen shot of a completed question.

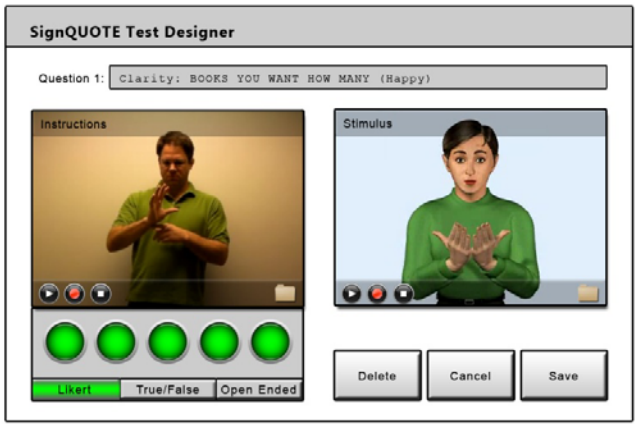

Figure 4: Question editor screen from TestDesigner module of SignQUOTE.

The resulting test design is stored in an XML-based configuration file which specifies the test's presentation in the TestServer component. Researchers administer the test by emailing invitations containing the hyperlink.

6.3 TestServer

All information and instructions in SignQUOTE's TestServer, from informed consent to post-test questionnaire, are presented in signed language. Figure 5 shows the screen layout for a closed-ended question using a Likert scale. Recordings of the test moderator appear in the upper right window and test stimuli appear on the left. The participant views instructions from the moderator and observes test stimuli. The participant can view a stimulus multiple times and then answer questions in the lower-right response area. Across the top of the screen is a progress indicator.

As is apparent in the figure, there are no text labels associated with the Likert choices. Instead, the interface takes advantage of a unique visual aspect of sign language called indexing [14]. Indexing occurs in ASL when a person refers to an object or another person in the environment, and involves pointing at the entity. The signed instructions in this tool use indexing to refer to the response choices. This is analogous to asking a hearing person to respond to the choices of a Likert scale. The absence of text on the buttons means that there are no pre-determined semantics associated with the buttons, which maximizes the flexibility of the interface.

The tool also provides for open-ended questions by capturing responses via the participant's webcam, as shown in Figure 6. The test moderator asks the participant to sign their response for the webcam. The response area changes to show a webcam control. The participant signs a response and clicks the control when done. Participants are comfortable with this due to their previous experience using webcams. Further, the informed consent in our studies stipulates that the recorded responses are only used for collecting aggregate data and are destroyed at the end of the study.

6.4 Technical Details

Both SignQUOTE components are Adobe Flex applications. They run on Apache, use PHP for data collection, and WowZa Media Server2 or Red 5 for video streaming. All of these are free or free for small numbers of simultaneous sessions. Necessary hardware includes an Internet connection and a webcam.

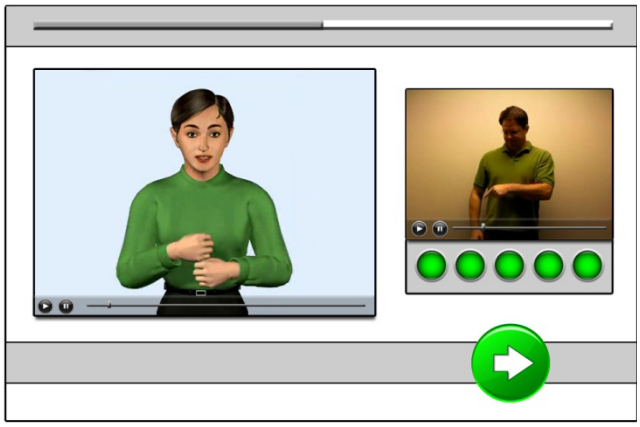

Figure 5: TestServer screenshot showing facilitator indexing the response buttons.

Figure 6: Interacting with SignQUOTE's TestServer.

Recorded video is stored in FLV format, and data are stored as text files. For the interpreter to view the videos, we set up a VLC player, and for our data analysis, we used MS Excel.

7. USABILITY

As discussed in [15], we conducted a usability study of the TestServer interface. This was an evaluation of TestServer's learnability, ease of navigation and functionality with a participant's choice of operating system and browser. This last was important because in practice, researchers will have no control over these choices. Five of the tests were conducted in a face-to-face setting and the remainder were conducted via Skype with a screen share. Participants were given a URL and told to visit the site. The URL contained a version of TestServer that presented an informed consent, and four animations depicting

signed sentences. Participants answered two open-ended and two closed-ended questions about each animation.

Paricipants were given no instructions on how to use the interface, but were encouraged to ask questions or give feedback about any difficulties they encountered. The facilitator observed participants as they navigated the site and noted their comments. After they completed the tasks using TestServer, they answered a few debriefing questions.

Participants uncovered issues when using TestServer on Apple operating systems, which has subsequently been addressed, and also suggested the addition of video controls to allow for scrubbing. The final debriefing questions pertained to the presentational techniques used in TestServer. Eighty-five percent of participants indicated that the indexing technique was very easily understood, and 100 percent agreed or strongly agreed with the statement "ASL is better than English for this type of test." Participants described the test approach as "inspired", "super-great", and "beneficial to the Deaf community".

8. CONFIDENTIALITY ISSUES

Maintaining confidentiality is paramount in studies involving human participants, which means that special measures need to be taken when recording a signed response to a question. For data analysis, we retain only an audio recording of the interpreter's voicing of the signed responses. Interpreters adhere to strict confidential and ethical standards set by Registry of Interpreters for the Deaf (RID) [16]. Destroying the video recordings is analogous to destroying recordings of face-to-face test sessions. In fact, since the researchers never need to see the faces of the participants, this method has an enhanced level of confidentiality.

Alternatively, it can be very valuable to retain the video recordings for study and analysis. Researchers wishing to retain the video would need to create an informed consent that explicitly asked for permission to do this.

9. COMPARING THE METHODS

Remote testing does have some inherent drawbacks, including the lack of control of the test environment as well as inability of a test facilitator to ask follow up questions based on participant responses. We wanted to see how SignQUOTE performed compared to face-to-face testing with an interpreter. To do this, we used SignQUOTE to test several animations that we had tested previously in a face-to-face setting [17].

As in the face-to-face test, we asked each participant to repeat the sentence and to rate its clarity on a 5-point Likert scale. Depending on the stimulus, we asked participants to rate the avatar's affect or to estimate the size of the object mentioned in the animation. This rating was also on a 5-point Likert scale. Finally, we asked the open-ended question, "Tell us how we can improve the animation." It is this last question that is essential to the goal of improved sign synthesis.

Twenty people participated in the face-to-face study and twenty-two participated in the remote study. The studies used identical stimuli and collected the same quantitative and qualitative data. As a first comparison we consider the quantitative results from the two studies. Of course, we are primarily interested in the ability to elicit qualitative feedback; however, it is helpful to check that the new test instrument is not overly skewing the quantitative data collected. As is commonly recognized, nonparametric tests such as Mann-Whitney are more appropriate than t-tests for analyzing this type of Likert data [18]. Figure 7 shows the medians and the two-tailed Mann-Whitney scores for the five stimuli in both studies.

Animation number	Face to Face Median	Remote Median	Mann-Whitney
1 (affect)	3	4	.38
2 (affect)	2	1	.21
3 (size)	5	4	.85
4 (size)	4	4	.30
5 (size)	5	5	.57

Figure 7: Results from the Size/Emotion Likert Scales

The Mann-Whitney metric attempts to measure the likelihood that a difference in distribution between two sample populations arises from random variation. A very low Mann-Whitney score (<.05) indicates that the differences in the two medians are statistically significant. As can be seen in Figure 7, the scores do not indicate that these differences are significant, thus it is quite possible that the difference between the face-to-face and remote medians resulted from randomness in the samples.

One of the possible contributing factors to the disparities between the face-to-face and remote results is the small sample size of the face-to-face test. Although face-to-face testing is the "gold standard" for eliciting qualitative feedback, it carries a high cost which often forces researchers to limit the number of participants.

In an effort to characterize the nature of the qualitative data, we created two metrics. The first, "elicitation" is the percentage of participants who gave substantive suggestions for improvement. Responses such as "It's fine," or "no comment" were omitted while responses such as "The brows should be up longer" or "She's signing too slow" were deemed substantive suggestions.

As seen in Figure 8, the differences between the face-to-face and remote testing methods were not statistically significant, but the percentages were consistently higher in the remote scenario. We do not know what caused this – it may be random variations or perhaps the absence of a human facilitator in the same room encouraged participants to offer suggestions more freely. Additional studies would be required to determine this.

	Face-to-face	Remote
Animation 1	50%	68.18%
Animation 2	65%	68.18%
Animation 3	35%	50%
Animation 4	55%	68.18%
Animation 5	40%	63.64%

Figure 8: Elicitation metric for eliciting qualitative data.

The first metric gave us a sense of the number of participants willing to give qualitative feedback, but we wanted to dig deeper. Our second metric, "overlap," was intended to give a sense of the scope of feedback in the remote data as compared with the face-to-face scenario. To compute overlap, we first created sets of distinct suggestions, one for the face-to-face data and one for the

remote data. We then calculated the intersection of the two sets and computed the following ratio:

$$\text{overlap} = \#(\text{f2f} \cap \text{remote}) / \#(\text{f2f})$$

An overlap of 100% would indicate that every suggestion occurring in the face-to-face set also occurred in the remote set. In Figure 9, the metric is expressed as k/p where k is the cardinality of the intersection and p is the cardinality of the face-to-face set.

Animation1	50%	(2 / 4)
Animation 2	40%	(2 / 5)
Animation 3	33%	(1 / 3)
Animation 4	50%	(3 / 6)
Animation 5	33%	(2 / 6)

Figure 9: Computing "overlap" of qualitative feedback gathered by remote testing.

The overlap scores indicate that each of the two scenarios elicited a number of unique suggestions. These findings are in line with results of previous studies that compared asynchronous remote testing with face-to-face testing of Web sites [19][20] and mobile applications [21]. In these studies, remote and face-to-face testing both uncovered some issues in common, but each technique also uncovered findings that were not surfaced by the other. Similar to the conclusion in these three studies, we believe remote testing is valuable as an addition to traditional face-to-face testing.

10. LESSONS LEARNED

We have been using TestServer for a little less than a year, and have adapted several of aspects of our testing protocol, including participant recruitment and the presentation of instructions and informed consent.

10.1 Recruiting participants

We found that the avenue that we used to recruit participants for face-to-face testing worked well with remote testing, which is to extend an invitation to participate on Deaf mailing lists. In face-to-face tests, we present a pre-test questionnaire in Sign to qualify participants, where we ask such questions as “Are you Deaf?”, “Are your parents Deaf?”, “How did you learn ASL?” and, “Did you go to a residential school or were you mainstreamed?” These questions are easily adaptable for use in SignQUOTE, where each question is presented as a video, and the response is an open-ended question.

Previously, the invitations were written in English, which is not the best approach because it introduces language barriers. In future, we plan to issue the invitations in ASL.

10.2 Instructions and Informed Consent

In the usability test, we found that it is important to keep individual video clips quite short (less than 45 seconds). If the clips are longer, people tend to lose focus. If the informed consent requires more time, it is better to subdivide the video into multiple segments.

10.3 A Complement not a Replacement

Remote testing does have several disadvantages, including lack of environmental control and no avenue for engaging in a discussion with a participant. A test facilitator cannot respond to participant feedback with follow up questions. For these reasons, we believe remote testing is a complement to face-to-face testing, not a replacement.

11. RESULTS

Using SignQUOTE significantly lowers the cost of conducting a test and researchers are not constrained by scheduling or geography when recruiting participants. As a result, tests can occur more often, and improvements to sign synthesis can occur more quickly.

Because it is configurable, and the testing interface is language-independent, SignQUOTE can be used with any sign language that uses indexing as a means of pronominalization. As well as testing synthesized sign language, the technology could serve as a platform for administering questionnaires. If researchers obtained permission from participants to archive the signed responses, SignQUOTE could also be used as a means to elicit signed exemplars for corpus building.

SignQUOTE is licensed under the GNU Affero General Public License. Both source and documentation are available for download at http://asl.cs.depaul.edu/signQUOTE.

12. ACKNOWLEDGEMENTS

Many thanks go to Jeff Karova for his valuable advice on the strategies for the effective use of ActionScipt. We would like to express our deep appreciation to the Deaf community for their continued support and participation, and to the superb interpreters, particularly Brianne DeKing, who have given us extremely valuable advice on best practices for promoting Deaf/hearing communication.

13. REFERENCES

[1] Hix, D. and Hartson, H. R. 1993. *Developing User Interfaces: Ensuring Usability through Product & Process.* New York, John Wiley and Sons.

[2] Ebling, M., and John, B. 2000. On the contributions of different empirical data in usability testing. *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (Brooklyn, NY, August 17-19, 2000) DIS'00 ACM New York, NY, 289 – 296 DOI=http://doi.acm.org/10.1145/347642.347766 .

[3] Waterson, S., Landay, J., and Matthews, T. 2002. In the lab and out in the wild: remote web usability testing for mobile devices. *Extended Abstracts on Human Factors in Computing systems* (Minneapolis, Minnesota, April 20 - 25, 2002) CHI EA ‘02. ACM, New York, NY, 796-797. DOI= http://doi.acm.org/ 10.1145/506443.506602.

[4] 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. *DePaul CTI Research Conference* (Chicago, IL, November 4, 2000). Available at http://asl.cs.depaul.edu/publications.html

[5] Kipp, M., Heloir, A. and Nguyen, Q. 2011 Sign Language Avatars: Animation and Comprehensibility. *Proceedings of the 11th International Conference on Intelligent Virtual Agents* (IVA-11), Springer. Available at http://embots.dfki.de/doc/Kippetal11.pdf

[6] Toro, J. A. 2005. *Automatic verb agreement in computer synthesized depictions of American Sign Language*. Doctoral Dissertation. UMI Order Number: UMI Order No. AAT 3175257, DePaul University.

[7] Valli, C., Lucas, C. and Mulrooney, K. 2005. *Linguistics of American Sign Language: An Introduction.* 4th ed. Gallaudet University Press, Washington, DC.

[8] Scholtz, J. 2001. Adaptation of traditional usability testing methods for remote testing. *Proc. Annual Hawaii International Conference on System Sciences* (Maui, HI, January 3-6, 2001), 8-15.

[9] Thompson, K., Rozanski, E., and Haake, A. 2004. Here, there, anywhere: Remote usability testing that works. *Proc. 5th Conference on Information Technology Education* (Salt Lake City, UT, October 28 - 30, 2004). CITC5 '04. ACM, New York, NY, 132-137. DOI=http://doi.acm.org/10.1145/1029533.1029567. .

[10] Duarte, K., Gibet, S., and Courty. 2011. Challenges and solutions for the SignCom data-driven signing avatar. Presented at the *First International Workshop on Sign Language Translation and Avatar Technology* (Berlin, Germany, January 10-11, 2001) SLTAT-2011.

[11] Hong, J., Heer, J., Waterson, S., and Landay, J. 2001. WebQuilt: A proxy-based approach to remote web usability testing. *ACM Transactions on Information Systems*, 19, 3, (July, 2001), 263-285. DOI=http://doi.acm.org/10.1145/502115.502118.

[12] Petrie, H., Hamilton, F., King, N., and Pavan, P. 2008. Remote Usability Evaluations with Disabled People. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (Montréal, Quebec, Canada, April 22-27, 2008). CHI'06 ACM, New York, NY, 1133-1141. DOI=http://doi.acm.org/10.1145/1124772.1124942.

[13] Hogg, N., Lomicky, C. and Weiner, S. 2008. Computer-Mediated Communication and the Gallaudet University Community: A Preliminary Report. *American Annals of the Deaf*, 153, 1, (Spring 2008), 89-96. .

Erting, E. 1992. Deafness & Literacy: Why Can't Sam Read? *Sign Language Studies*, 75, (Summer, 1992), 97-112.

[14] Baker-Shenk, C. and Cokely, D. 1980. *American Sign Language: A Teacher's Resource Text on Grammar and Culture*. Gallaudet University Press, Washington, DC.

[15] Schnepp, J., and Shiver, B. 2011. *A Deaf-Accessible Tool for Remote Usability Testing.* Submitted to Assets 2011.

[16] Registry of Interpreters for the Deaf. 2011. *NAD-RID Code of Professional Conduct,* Available at www.rid.org/UserFiles/File/NAD_RID_ETHICS.pdf

[17] 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: Corpora and Sign Language Technologies* (Valetta, Malta. May 23, 2010) LREC 2010. 217-220. Available at http://www.sign-lang.uni-hamburg.de/lrec2010/lrec-cslt-01.pdf .

[18] Gregoire, T. G. and Driver, B. L., 1987. Analysis of ordinal data to detect population differences. *Psychological Bulletin*, 101, 1, (January 1987), 159-165.

[19] Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, 2002. M. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. *Usability Professionals Association Conference* (Orlando, Florida, July 8-12, 2002) 32.

[20] Andreasen, M., Nielsen, H., Schrøder, Stage, Jan. 2007. What Happened to Remote Usability Testing? An Empirical Study of Three Methods. *Proceedings of the 25th international conference on Human factors in computing systems* (San Jose, CA, April 28 – May 3 2007) CHI'07 ACM, New York, NY, 1405-1414. DOI= http://doi.acm.org/ 10.1145/1240624.1240838

[21] Bruun, A., Gull, P., Hofmeister, L., Stage, J. 2009. Let your users do the testing: a comparison of usability testing methods. *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (Boston, MA, April 4-9, 2009) CHI'09 ACM, New York, NY,1619-1628. DOI=http://doi.acm.org/10.1145/1518701.1518948