This is a copy of the manuscript accepted to IEEE Computer Graphics and Applications on June 21, 2015.

The final version is available at the following link:
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7274434&isnumber=7274393
doi: 10.1109/MCG.2015.94

Toward Standard Usability Questionnaires for Handheld Augmented Reality

Marc Ericson C. Santos, Jarkko Polvi, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor, Hirokazu Kato

Handheld devices such as smartphones and tablet computers now have powerful processors, large screens, and built-in location sensors and cameras. These features of handheld devices make them convenient platforms for augmented reality (AR) – the seamless integration of virtual objects to real environments. Handheld augmented reality (HAR) affords many new ways of interacting with digital content. It is finding applications in various industries such as entertainment, marketing and sales, education and training, navigation and tourism, and social networking. Although several applications have been adopted by general consumers, HAR remains limited and researchers are continuously developing more intuitive interactions using handheld devices.

Usability refers to how well target users can use a functionality of a system [1] to accomplish a specific task. Usability studies are important for assessing and iteratively improving AR systems [2]. Among the widely used evaluation technique in user studies are subjective measurements such as questionnaires, user ratings, or judgments. For AR systems, researchers have used the System Usability Scale and the NASA Task Load Index for quantifying general system usability and workload, respectively. For handheld devices, the Mobile Phone Usability Questionnaire (MPUQ) enumerates the various questionnaires for common uses of mobile phones [3]. These questionnaires have been previously evaluated and studies support their validity and reliability. However, these standard questionnaires do not consider specific perceptual and ergonomic issues common to HAR systems. As such, researchers complement these evaluation tools with their own questionnaires. These questionnaires are not always tested for validity and reliability. Moreover, the questions tend to be specific to the features of their HAR system.

In response to the lack of valid and reliable evaluation tools for HAR, we developed the HAR Usability Scale (HARUS) which is composed of two sub-questionnaires, namely the comprehensibility scale and the manipulability scale. We designed the questionnaires based on a review of usability evaluations of HAR systems. We then evaluated the validity and reliability of HARUS in four experiments, and discuss some insights gathered from using HARUS in our own user studies. Researchers and professionals involved in developing HAR applications can directly use our questionnaire to evaluate their own HAR applications, or modify it with considerations of the insights presented in this paper.

Designing the Questionnaires

To create the HARUS, we followed a five-step method for developing and testing questionnaires [4]. The first three steps correspond to designing the questionnaire by studying the background, then conceptualizing the questionnaire, and finally deciding the format and data analysis. We conducted a systematic literature review of research papers to list the HAR issues that are either raised by users and expert reviewers, or observed by the experimenters when using various HAR systems [5]. We classified the issue as either a perceptual issue or an ergonomic issue as shown in Table 1. Some of the perceptual and ergonomic issues arise from the mobility of HAR. For example, a user who walks into a location with bad signal reception would feel that the HAR display is slow and unresponsive. This lessens the comprehensibility of the HAR.

Given the two types of issues, the goal of design for HAR is to have no perceptual issues and no ergonomic issues. We refer to these qualities as comprehensible and manipulable, respectively. Comprehensibility is the ease of understanding the information presented by the HAR system, whereas manipulability is the ease of handling the HAR system as the user performs the task. In our questionnaire concept, a perfect HAR system (for a target user group in a particular task) would score 100 on measures of comprehensibility and manipulability. Thus, we are approximating HAR usability by just considering these two factors.

We patterned the questionnaire format and data analysis from the System Usability Scale (SUS) [6] and design rules prescribed by Floyd J. Fowler Jr. and Carol Cosenza [7]. The questionnaire is composed of statements corresponding to the issue listed in Table 1. These statements break down comprehensibility and manipulability into multiple measures to which users can respond to. One particular question may appear to be similar with the other questions within the set of eight because each question acts as a measure to the same construct. However, these set of statements is not exhaustive operationism of manipulability and comprehensibility. Rather, they are measures belonging to an extensible set of indicators for these two defined constructs.

Table 1. Specific HAR Issues Encountered and HARUS Statements

Specific Issues Encountered	**HARUS Statements**
Perceptual Issues: • The tracking is unstable due to the ambient light or bad sensor fusion. • The virtual objects are not well-registered. • The application is lagging or has intolerable latency. • The content was excessive and/or has poor quality. • The display induces too much cognitive load. • The download time of the content is too slow. • The screen is not legible due to outdoor ambient light. • The screen is not legible due to reflection or glare. • Depth is not understood or underestimated.	Comprehensibility Statements: 1. I think that interacting with this application requires a lot of mental effort. 2. I thought the amount of information displayed on screen was appropriate. 3. I thought that the information displayed on screen was difficult to read. 4. I felt that the information display was responding fast enough. 5. I thought that the information displayed on screen was confusing. 6. I thought the words and symbols on screen were easy to read. 7. I felt that the display was flickering too much. 8. I thought that the information displayed on screen was consistent.
Ergonomic Issues: • The application causes fatigue after extended use. • The device is too bulky or too heavy. • Hand interactions are difficult to perform. • The application is not responsive or provides no feedback. • The keypad is too small.	Manipulability Statements: 9. I think that interacting with this application requires a lot of body muscle effort. 10. I felt that using the application was comfortable for my arms and hands. 11. I found the device difficult to hold while operating the application. 12. I found it easy to input information through the application. 13. I felt that my arm or hand became tired after using the application. 14. I think the application is easy to control. 15. I felt that I was losing grip and dropping the device at some point. 16. I think the operation of this application is simple and uncomplicated.

Related Work on Usability Issues in HAR

David Drascic and Paul Milgram were first to discuss the perceptual issues in augmented reality.[1] A study by J. Edward Swan II and colleagues distinguishes perceptual issues from other ergonomic issues of mobile augmented reality systems.[2] Ernst Kruijff and colleagues summarized these perceptual issues by considering the human visual processing system and interpretation pipeline.[3] In another study, Eduardo Veas and Ernst Kruijff summarized ergonomic issues in handheld augmented reality and evaluated several handheld devices to address these ergonomic issues.[4]

References

1. D. Drascic and P. Milgram, "Perceptual Issues in Augmented Reality," Proc. SPIE Vol. 2653: Stereoscopic Displays and Virtual Systems III, 1996, pp. 123–134.
2. J. E. II Swan and J. L. Gabbard, "Perceptual and Ergonomic Issues in Mobile Augmented Reality for Urban Operations" Naval Research Laboratory Technical Memorandum Report, 2003.
3. E. Kruijff, J. E. II Swan and S. Feiner, "Perceptual Issues in Augmented Reality Revisited," Proc. ISMAR, 2010, pp. 3–12.
4. E. Veas and E. Kruijff, "Vesp'r: Design and Evaluation of a Handheld AR Device," Proc. ISMAR, 2008, pp.43–52.

HARUS is composed of 16 statements listed in Table 1. The first eight statements comprise the comprehensibility scale, and the last eight statements comprise the manipulability scale. These questionnaires are answerable by asking users how much they agree or disagree to the statements. For our experiments, we used a seven-point Likert scale with the first point labeled as "Strongly Disagree" and seventh point labeled as "Strongly Agree." To compute the comprehensibility and manipulability scores, we first convert the scores to range from zero to six. For the positively-stated items, we subtract one from the user response. For the negatively-stated items, we subtract the user response from seven. We add all these responses, and map the sum to a range of 0 to 100. To do this, we divide the sum by the highest possible score of 48 and then multiply the result by 100. Finally, we obtain the HARUS score by getting the average of the comprehensibility and manipulability scores.

Testing Questionnaires in Experiments

The last two steps in the five-step method for developing and testing questionnaires are conducting tests of validity and reliability. We gathered some evidences of the validity and reliability of HARUS in four experiments summarized in Table 2. We validated HARUS by showing concurrent validity, a kind of criterion-oriented validation procedure. Validity is a matter of degree, not all or nothing. In our experiments, we measured the degree of correlation between HARUS and other acceptable measures of usability.

Our experiment scenarios involve simple but relevant HAR tasks. In our experiments, we used graphics symbols such as arrows for pointing to specific three-dimensional locations in the real world. Then, accompanying text and sprite animations communicate information relevant to the task. Experiments 1 and 4 are basic content authoring tasks with experiment 4 testing the fundamental authoring task of positioning a virtual object. Experiments 2 and 3 are content consumption tasks wherein the user examines content in order to accomplish the task.

We compared HARUS with objective measures of usability namely, time on task, number of words, study duration and positioning error. We also compared HARUS to subjective measures namely, the SUS, the MPUQ, and the Instructional Material Motivation Survey (IMMS) [8]. Intuitively, we know that HARUS would correlate with other measures of usability because of its design. However, it is interesting to know the strength of the correlation in actual experiments. This strength represents the degree of validity of HARUS.

For the last step, we measured the reliability or the precision of HARUS by computing the Cronbach's alpha. The Cronbach's alpha is a measure of the internal consistency of a questionnaire [9].

Table 2. Summary of Experimental Design

Task	Experimental Platform	Participants	Usability Metrics
Annotating Text: Register feature points then add virtual text on objects	Simple Authoring App iPad 2, iOS6 PointCloud SDK	n=18 22 to 41 years old (M=27, SD=4)	SUS Time on task
Status Reporting: Create a report on the status of devices found in a room	Simple Viewing App iPad Mini, iOS7 PointCluod SDK	n=20 19 to 46 years old (M=28, SD=8)	SUS MPUQ Number of words
Memorizing Words: Memorize 30 words in a refreshment area for one week	FlipPin App iPad 2, iOS6 ARToolKit	n=18 23 to 32 years old (M=26, SD=3)	SUS IMMS Study duration
Positioning Arrows: Place virtual arrows on target locations by adjusting it to the correct place	Device-centric Method iPad 4, iOS7 PointCloud SDK	n=23 22 to 42 years old (M=30, SD=5)	SUS Time on task Positioning error

Experiment 1: Annotating Text

HAR systems can be used for creating digital content in situ. In this experiment, users evaluated a HAR system for annotating text on real objects found in the environment. We implemented a simple HAR authoring tool for annotating text on real objects as shown in Figure 1 (right-most). We used the PointCloud SDK to detect some natural feature points in the environment. To register the feature points, the user must move the iPad 2 from side-to-side (Figure 1, left-most). Once the system detects enough feature points, the user can add a text label on the scene (Figure 1, middle).

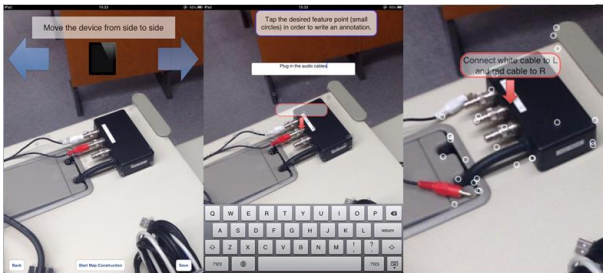

Figure 1. Simple HAR Authoring Tool for Annotating Text

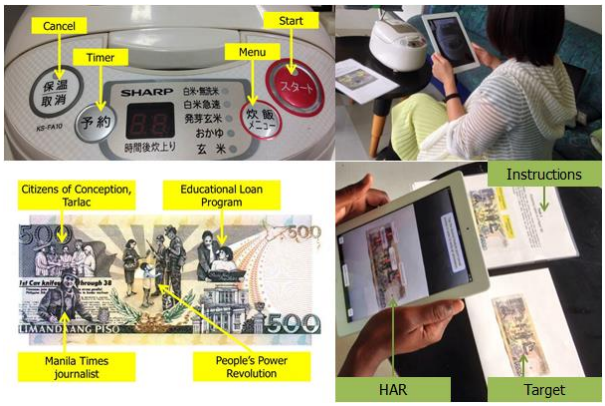

Figure 2. Authoring Tasks

We asked our user study participants to annotate English translations on a rice cooker and trivia on a paper bill as shown in Figure 2. We did not give a time limit to do the tasks and the participants were free to opt out at any time. We offered this option because we found out in a pilot study that some people fail to do the registration procedure. After finishing the task or giving up, the participants answered the SUS and the HARUS questionnaires. Half of the participants answered the SUS first, whereas half answered the HARUS first. We evaluated HARUS by comparing it to the SUS and task on time with the following hypotheses:

H1. HARUS and SUS have a positive relationship.
H2. HARUS and time on task have a negative relationship.

The participants gave the HAR authoring tool an average of SUS score of 62 (SD=22) and an average HARUS score of 65 (SD=16), and they finished the task with an average time on task of 8.1 minutes (SD=2.5). The HARUS score has a very strong positive relationship with the SUS score and a strong negative relationship with time on task as shown in Table 3. Moreover, both comprehensibility and manipulability scores have very strong positive relationship with the SUS score, and a strong negative relationship with time on task as shown in Table 4.

Table 3. Correlations (r) of HARUS, SUS and Time on Task

	1	2	3
1. HARUS	1.00		
2. SUS	0.87***	1.00	
3. Time on Task	-0.51*	-0.58**	1.00

*significant at 0.05 level
**significant at 0.01 level
***significant at 0.001 level

Table 4. Correlations (r) of Comprehensibility, Manipulability, SUS and Time on Task

	1	2	3	4
1. Comprehensibility	1.00			
2. Manipulability	0.40	1.00		
3. SUS	0.75***	0.72***	1.00	
4. Time on Task	-0.45*	-0.41*	-0.59**	1.00

*significant at 0.05 level
**significant at 0.01 level
***significant at 0.001 level

Experiment 2: Status Reporting

For this experiment, users evaluated an application for viewing virtual notes on real objects. HAR applications commonly require users to read information associated with the real environments. Examples include advertisement, scientific information, etc. We implemented a HAR application that enables users to view text annotations on real objects as shown in Figure 3.

Figure 3. HAR for Viewing Annotations on Equipment

The participants assumed the role of a newly-hired equipment maintenance staff. As their first job, they need to write a report on the status of equipment by viewing annotations made by the previous maintenance staff. They then filled a report form with three columns: device, issue and recommended action. To make the report, the participants need to gather information from the HAR and the devices such as model, serial numbers, brand, etc. This kind of work-support task is not limited to AR that uses head-mounted displays. Some researchers recommend the use of HAR because it is less intimidating for workers and they find it easier to share with their co-workers [10].

We gave the participants a time limit of 15 minutes to finish the task. After the task, we asked them to answer three questionnaires, namely HARUS, SUS and eight questions from the MPUQ's affective aspects and media properties (AAMP). We evaluated HARUS by comparing it to the SUS, MPUQ and number of words written on the report with the following hypotheses:

H3. HARUS and SUS have a positive relationship.
H4. HARUS and MPUQ have a positive relationship.
H5. HARUS and number of words have a positive relationship.

We observed that the natural interaction pattern is as follows: First, the participants find a suitable angle that would reveal the virtual information. They then freeze the screen and settle to a more relaxed pose. Lastly, they switch between reading the screen and inspecting the device when filling the report form.

Only one participant was not able to finish the report under 15 minutes. The rest were able to finish the report with an average time of 9.9 minutes (SD=1.9). The participants made reports consisting of an average of 73.5 words (SD=19.5) about 13 individual devices. They gave the HAR an average SUS score of 80 (SD=11), whereas the average HARUS and AMMP scores were 74 (SD=13) and 80 (SD=13), respectively.

The HARUS scores have a very strong positive relationship with the SUS and MPUQ scores as shown in Table 5. Moreover, both comprehensibility and manipulability scores have strong positive relationship with the SUS and MPUQ as shown in Table 6. However, comprehensibility has a stronger correlation with the SUS and the MPUQ probably because of the nature of the task which focuses on reading some virtual information. This is in contrast to the authoring task in Experiment 1 which required moving the iPad from side-to-side to register feature points.

We did not find any significant relationship between HARUS and number of words probably because low word count could mean either lacking in information (bad quality) or simply concise (good quality). However, a strong positive relationship exists between the manipulability score and number of words. In other words, people who found the HAR easy to handle tend to write more words on their report. We find this logical and we believe that there are trade-offs in user performance for activities that split the use of hands (e.g. handling the HAR and hand-writing a report).

Table 5. Correlations (r) of HARUS, SUS, MPUQ and Number of Words

	1	2	3	4
1. HARUS	1.00			
2. SUS	0.79***	1.00		
3. MPUQ (AAMP)	0.75***	0.82***	1.00	
4. Number of Words	0.12	-0.23	0.43*	1.00

*significant at 0.05 level
***significant at 0.001 level

Table 6. Correlations (r) of Comprehensibility, Manipulability, SUS, MPUQ and Number of Words

	1	2	3	4
1. Comprehensibility	1.00			
2. Manipulability	0.34	1.00		
3. SUS	0.70***	0.58**	1.00	
4. MPUQ (AAMP)	0.68***	0.54*	0.82***	1.00
5. Number of Words	-0.19	0.41*	0.23	0.43*

*significant at 0.05 level
**significant at 0.01 level
***significant at 0.001 level

Experiment 3: Memorizing Words

HAR can be used to support learning in natural environments. It transforms the real environment into a learning experience by adding a layer of virtual content. In this experiment, we evaluated a HAR system for memorizing Filipino vocabulary inside a refreshment area. We implemented a simple application for displaying text, audio and sprite sheet animation on top of real objects. We used the ARToolKit to measure the pose of the iPad with respect to the target object. Figure 4 shows screenshots of the HAR system.

We tagged 30 objects found in a refreshment area as shown in Figure 5. Each of the 30 real objects is associated with a Filipino word (15 nouns, 15 verbs). We decided to use Filipino as the target language to minimize the effects of proficiency in their first language. Objects used to teach Filipino nouns are annotated with the word itself as virtual labels. Those teaching Filipino verbs are annotated with sprite sheets that demonstrating the action.

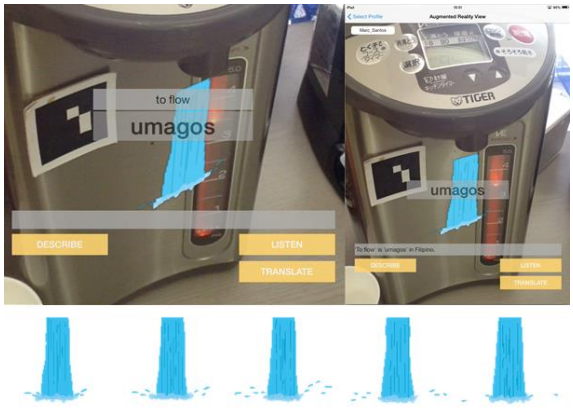

Figure 4. HAR for Word Memorization in a Real Environment. The sprite sheet animation illustrates the word "umagos" which is the Filipino word for "to flow."

Figure 5. Memorizing Words in a Refreshment Area

The participants studied Filipino for five days with a recommended study time of 15 minutes a day. However, they can use the application as much as they want. Each of the Filipino words have two to three descriptions of the scene that can be accessed by pressing the describe button. Each plays one sound file of the proper pronunciation (listen button) and presents one translation (translate button). Each participant has a user account which we monitored by logging activities for each account. On the last day of using the system, we asked the participants to answer three questionnaires, namely HARUS, SUS and IMMS. We evaluated HARUS by comparing it to the SUS, IMMS and the total study time or duration with the following hypotheses:

H6. HARUS and SUS have a positive relationship.
H7. HARUS and IMMS have a positive relationship.
H8. HARUS and study duration have a positive relationship.

The participants studied for an average of 42.7 minutes (SD=19.5) for five days. On the average, the participants gave the application an SUS score of 74 (SD=12). However, they rated the application 61 (SD=15) on the HARUS. This difference is the largest that we observed between the SUS and HARUS scores. Lastly, the participants gave the HAR an average IMMS score of 59 (SD=14). The HARUS has a strong positive relationship with the SUS, the IMMS, and the study time as shown in Table 7.

The participants who gave higher HARUS scores tend to find the interface more motivating. They also tend to study longer with the interface. All of these correlations are significant and support hypotheses 6 to 8. Moreover, comprehensibility and manipulability have a strong positive relationship (except between manipulability and study duration) as shown in Table 8. Similar to Experiment 2, we found that comprehensibility has a stronger relationship to the SUS score and study duration probably because of the nature of the task which focuses on memorization.

Table 7. Correlations (r) of HARUS, SUS, IMMS and Study Duration

	1	2	3	4
1. HARUS	1.00			
2. SUS	0.68***	1.00		
3. IMMS	0.61**	0.55**	1.00	
4. Study Duration	0.42*	0.49*	0.13	1.00

*significant at 0.05 level
**significant at 0.01 level
***significant at 0.001 level

Table 8. Correlations (r) of Comprehensibility, Manipulability, SUS, IMMS and Study Duration

	1	2	3	4
1. Comprehensibility	1.00			
2. Manipulability	0.60**	1.00		
3. SUS	0.76***	0.47*	1.00	
4. IMMS	0.49*	0.60**	0.55**	1.00
5. Study Duration	0.53*	0.24	0.49*	0.13

*significant at 0.05 level
**significant at 0.01 level
***significant at 0.001 level

Experiment 4: Positioning Arrows

Positioning virtual objects is one of the most important tasks in authoring AR contents. Currently, HAR has no established interaction metaphors thus various methods of doing specific tasks need to be evaluated. One such task is adjusting the 3D position of a virtual object in the real environment. For this experiment, we implemented a device-centric method similar to the work of Anders Henrysson and colleagues [11]. When the user selects the virtual object on the screen, the position of the virtual object becomes fixed relative to the movement of the device. As such, the user can drag the virtual object by moving the handheld device in any direction.

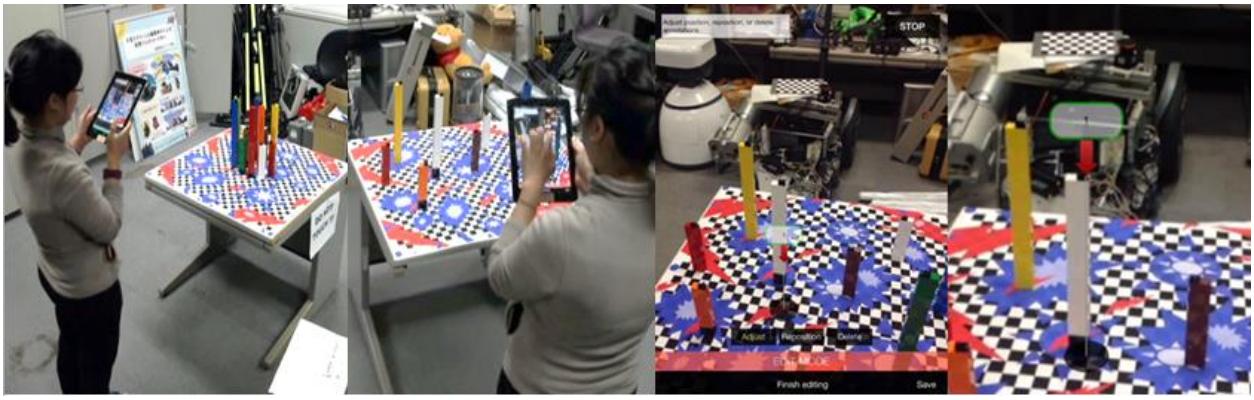

Figure 6. Adjusting arrows to target pillars with different heights

We asked the participants to position arrows on top of each of the eight pillars in Figure 6. They did this twice, once with the pillars concentrated in the center (Figure 6, left most) and once with the pillars farther apart (Figure 6, second from left). After the task, we asked all 23 of them to answer HARUS. Of the 23, we only asked 7 to answer the SUS to save time. We took note of the time on task and we measured the distance of each arrow to the target pillar as the positioning error. In this experiment, we focused on comparing HARUS with time on task and total positioning error with the following hypotheses:

H9. HARUS and SUS have a positive relationship.
H10. HARUS and time on task have a negative relationship.
H11. HARUS and positioning error have a positive relationship.

The participants spent an average time of 18.2 minutes (SD=7.8) on the task with an average total positioning error of 22.6 mm (SD=9.3). On the average, the participants gave the application an SUS score of 57 (n=7, SD=19) and they rated the application 58 (SD=15) on the HARUS. Based on seven participants, the HARUS, comprehensibility and manipulability have very strong positive relationships with the SUS (Table 9). Based on 23 participants, the HARUS, comprehensibility and manipulability have strong negative

relationships with the total amount of error but not with time on task. In other words, participants who were less accurate with positioning the arrows tend to give lower usability scores to the application. For this experiment, we found evidences supporting hypotheses 9 and 11, but not 10.

Table 9. Correlations (r) of HARUS, Comprehensibility, Manipulability and SUS

	1	2	3	4
1. HARUS	1.00			
2. Comprehensibility	0.91**	1.00		
3. Manipulability	0.93**	0.68	1.00	
4. SUS	0.90**	0.84*	0.83*	1.00

*significant at 0.05 level
**significant at 0.01 level

Table 10. Correlations (r) of HARUS, Time on Task and Total Positioning Error

	1	2	3
1. HARUS	1.00		
2. Time on Task	-0.06	1.00	
3. Positioning Error	-0.63**	-0.21	1.00

**significant at 0.01 level

Table 11. Correlations (r) of Comprehensibility, Manipulability, Time on Task and Total Positioning Error

	1	2	3	4
1. Comprehensibility	1.00			
2. Manipulability	0.61**	1.00		
3. Time on Task	-0.03	-0.07	1.00	
4. Positioning Error	-0.49*	-0.62*	-0.21	1.00

*significant at 0.05 level
**significant at 0.01 level

Reliability of HARUS in Four Experiments

We computed the Cronbach's alpha in all four experiments to measure the internal consistency of HARUS, the comprehensibility scale and the manipulability scale. Table 12 lists the alphas for the four experiments. All of the alphas obtained are between 0.7–0.9 thus HARUS and its sub-questionnaires have good internal consistency.

Table 12. Cronbach's Alpha in Four Experiments

	HARUS	Comprehensibility	Manipulability
Annotating Text	0.83	0.74	0.71
Status Reporting	0.83	0.80	0.81
Memorizing Words	0.87	0.79	0.83
Adjusting Arrows	0.88	0.82	0.81

Summary of Findings

We designed a new usability scale for HAR systems and we conducted four experiments to evaluate its validity and reliability. We observed the following in our experiments:

1. HARUS and SUS have a very strong relationship in all four experiments.
2. In experiment 1, the HARUS score increases as the time taken to finish the task decreases.
3. In experiment 3, the participants who gave higher HARUS scores tend to study for longer periods of time.
4. In experiment 4, the participants who gave a higher HARUS score tend to be more accurate at positioning the virtual objects.
5. In experiments 2 and 3, the HARUS score increases with self-report measures of positive emotions and motivation, respectively.

6. In experiment 3, we observed a large margin between the HARUS and SUS scores. We gained a good SUS score (above 70 [12]) whereas the HARUS score is much lower. This may be a situation when the SUS does not capture the problems unique to HAR and should be further investigated through user studies.
7. HAR usability, comprehensibility and manipulability scales demonstrated good internal consistency in all our experiments.
8. The manipulability and comprehensibility scales have varying degrees of relationship strength with the SUS, time on task, study time, total positioning error, positive emotions and motivation. These separate scales provide more insight when analyzing HAR.
9. In experiments 1 and 2, comprehensibility and manipulability only correlate moderately. As such, these two constructs should be analyzed independently from each other because it is possible for a HAR system to suffer more from perceptual issues than ergonomic issues, or vice versa.

Conclusions

HAR is novel interface that has a high potential for becoming a mainstream technology. It is useful for delivering various content in many fields of application. The development of new interaction metaphors and HAR systems must also be accompanied with the development of new evaluation tools and frameworks. Valid and reliable questionnaires are important for conducting user studies to iteratively improve HAR interaction metaphors and HAR systems.

We designed HARUS with sub-questionnaires (comprehensibility scale and manipulability scale) based on ergonomic and perceptual issues of HAR. This approach is advantageous because there are cases wherein standard questionnaires like the SUS do not capture the unique issues in HAR. Moreover, distinguishing between perceptual and ergonomic issues reveals that comprehensibility and manipulability are separate constructs. The usability of a system can suffer more from one of these two separate constructs, and efforts in improving on one could significantly improve the whole system.

We showed the validity and reliability of HARUS in four experiments. Our experimental scenarios arise from our own interest in using HAR for displaying information related to nearby scenes and objects. Our experiments are not exhaustive of various HAR scenarios. In particular, we do not have experiments of HAR systems for the far-field and those applied outdoors. As such, it would be interesting to see if HARUS is also valid and reliable for these specific cases wherein HAR is commonly applied. In addition, we recommend evaluating HAR systems that features more graphic symbols that are three-dimensionally registered to the environment.

HARUS is a tool for evaluating HAR applications with users as they perform specific tasks. HARUS aggregates usability, comprehensibility and manipulability into single scores. This score can be used by researchers and professionals to compare between iterations of the same system, to prioritize among several features of the same system, and to benchmark against previously evaluated implementations of a HAR system.

We plan to continue using HARUS in our own user studies to evaluate our HAR systems. We encourage other HAR researchers and professional to use these questionnaires, or apply the questionnaire design framework we discussed. Although we learned a lot from four experiments, it is not enough. HARUS needs to be compared with other objective measures of usability such as the use of wearable biosensors and built-in sensors of the handheld devices like accelerometers and gyroscopes.

Acknowledgement
This work was supported by the Grant-in-Aid for JSPS Fellows, Grant Number 15J10186.

References:

[1] J. Nielsen, Usability Engineering, Elsevier, 1994.

[2] J. Gabbard and J. Swan, "Usability Engineering for Augmented Reality: Employing User-Based Studies to Inform Design," *IEEE Transactions on Visualization and Computer Graphics,* vol. 14, no. 3, pp. 513-525, May 2008.

[3] Y. S. Ryu and T. L. Smith-Jackson, "Reliability and validity of the mobile phone usability questionnaire (MPUQ)," *Journal of Usability Studies,* vol. 2, no. 1, pp. 39-53, 2006.

[4] R. B. Radhakrishna, "Tips for developing and testing questionnaires/instruments," *Journal of Extension,* vol. 45, pp. 1-4, 2007.

[5] M. E. C. Santos, J. Polvi, T. Taketomi, G. Yamamoto, C. Sandor and H. Kato, "A Usability Scale for Handheld Augmented Reality," in *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, 2014.

[6] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," in *Human Centered Design*, Springer, 2009, pp. 94-103.

[7] F. J. Fowler and C. Cosenza, "International Handbook of Survey Methodology," E. D. de Leeuw, J. J. Hox and D. A. Dillman, Eds., Taylor & Francis, 2008, pp. 136-159.

[8] W. Huang, W. Huang, H. Diefes-Dux and P. K. Imbrie, "A preliminary validation of Attention, Relevance, Confidence and Satisfaction Model-based Instructional Material Motivational Survey in a Computer-based Tutorial Setting," *British Journal of Educational Technology,* vol. 37, no. 2, pp. 243-259, 2006.

[9] J. A. Krosnick and S. Presser, "Handbook of Survey Research," P. V. Marsden and J. D. Wright, Eds., Emerald Group Publishing Limited, 2010, pp. 263-313.

[10] G. Schall, E. Mendez, E. Kruijff, E. Veas, S. Junghanns, B. Reitinger and D. Schmalstieg, "Handheld Augmented Reality for underground infrastructure visualization," *Personal and Ubiquitous Computing,* vol. 13, no. 4, pp. 281-291, 2009.

[11] A. Henrysson, M. Billinghurst and M. Ollila, "Virtual Object Manipulation Using a Mobile Phone," in *Proceedings of the International Conference on Augmented Tele-existence*, 2005.

[12] A. Bangor, P. T. Kortum and J. T. Miller, "An Empirical Evaluation of the System Usability Scale," *International Journal of Human-Computer Interaction,* vol. 24, no. 6, pp. 574-594, 2008.

Marc Ericson C. Santos is a doctoral student at Nara Institute of Science and Technology. He received his B. S. and M. S. from Ateneo de Manila University in the Philippines. His research interests include human-computer interaction and augmented reality. Contact him at chavez-s@is.naist.jp.

Jarkko Polvi is a doctoral student at the Nara Institute of Science and Technology. He received his B. S. and M. S. from the University of Oulu. His research interests include human-computer interaction and mobile augmented reality systems. Contact him at jarkko-p@is.naist.jp.

Takafumi Taketomi is an assistant professor at Nara Institute of Science and Technology. He received his M. E. and Dr. Eng. from Nara Institute of Science and Technology. His main research work is in computer vision and human-computer interaction. Contact him at takafumi-t@is.naist.jp.

Goshiro Yamamoto is an assistant professor at Nara Institute of Science and Technology. He received his B. E., M. E., and Dr. Eng. from Osaka University. His interests are in human-computer interaction, projection-based augmented reality, and wearable computing systems. He was a JSPS research fellow from 2008 to 2009. Contact him at goshiro@is.naist.jp.

Christian Sandor is an associate professor at Nara Institute of Science and Technology, where he co-directs the Interactive Media Design Laboratory. His main research interest is augmented reality. In 2014, he received a Google faculty award for developing an augmented reality x-ray system for Google Glass. Contact him at sandor@is.naist.jp.

Hirokazu Kato is a professor at Nara Institute of Science and Technology, where he co-directs the Interactive Media Design Laboratory. His main research topic is augmented reality. He received the Virtual Reality Technical Achievement Award from IEEE VGTC in 2009, and the Lasting Impact Award from IEEE ISMAR in 2012. Contact him at kato@is.naist.jp.