A Usability Scale for Handheld Augmented Reality

Marc Ericson C. Santos[*]
Nara Institute of Science and Technology
Jarkko Polvi[†]
Nara Institute of Science and Technology
Takafumi Taketomi[‡]
Nara Institute of Science and Technology
Goshiro Yamamoto[§]
Nara Institute of Science and Technology
Christian Sandor[¶]
Nara Institute of Science and Technology
Hirokazu Kato[‖]
Nara Institute of Science and Technology

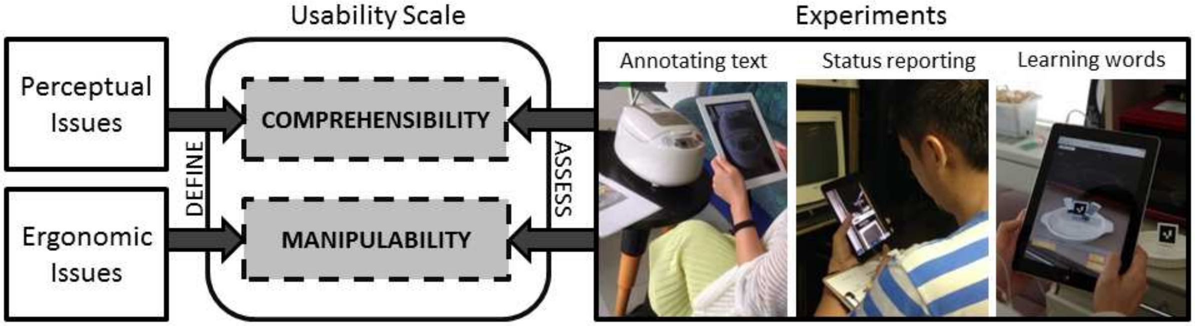

Figure 1: *We created a usability scale for evaluating handheld augmented reality (HAR) applications. We defined our usability scale based on perceptual and ergonomic issues encountered by users, and then assessed it using three experiments representative of common HAR tasks.*

Abstract

Handheld augmented reality (HAR) applications must be carefully designed and improved based on user feedback to sustain commercial use. However, no standard questionnaire considers perceptual and ergonomic issues found in HAR. We address this issue by creating a HAR Usability Scale (HARUS).

To create HARUS, we performed a systematic literature review to enumerate user-reported issues in HAR applications. Based on these issues, we created a questionnaire measuring *manipulability* – the ease of handling the HAR system, and *comprehensibility* – the ease of understanding the information presented by HAR. We then provide evidences of validity and reliability of the HARUS questionnaire by applying it to three experiments. The results show that HARUS consistently correlates with other subjective and objective measures of usability, thereby supporting its concurrent validity. Moreover, HARUS obtained a good Cronbach's alpha in all three experiments, thereby demonstrating internally consistency.

HARUS, as well as its decomposition into individual manipulability and comprehensibility scores, are evaluation tools that researchers and professionals can use to analyze their HAR applications. By providing such a tool, they can gain quality feedback from users to improve their HAR applications towards commercial success.

CR Categories: I.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; Evaluation/methodology I.5.2 [Information Interfaces and Presentation]: User Interfaces—Ergonomics; Evaluation/methodology

Keywords: augmented reality, evaluation method, handheld devices, usability, user studies

1 Introduction

Handheld augmented reality (HAR) enables consumer applications in entertainment, marketing and sales, education and training, navigation and tourism, and social networking [Gervautz and Schmalstieg 2012]. Quick but reliable evaluation tools are needed to ensure the commercial success of HAR applications. Evaluations are necessary to assess how well these applications address the needs of users, especially sensitive user groups such as children, laborers, and the elderly. Moreover, user feedback from evaluations is important to optimize revenues from these applications.

Usability, or the ease of using an interface [Nielsen 1994], is an important consideration that affects user adoption and user experience [Pilke 2004]. As such, researchers recommend usability evaluations that use cost-effective methods [Hix et al. 2004] and employ user-based studies [Gabbard and Swan 2008] to iteratively improve novel systems. Among the widely used evaluation techniques in augmented reality (AR) are subjective measurements such as questionnaires, user ratings, or judgements [Dünser et al. 2008].

[*] e-mail:chavez-s@is.naist.jp
[†] e-mail:jarkko-p@is.naist.jp
[‡] e-mail:takafumi-t@is.naist.jp
[§] e-mail:goshiro@is.naist.jp
[¶] e-mail:sandor@is.naist.jp
[‖] e-mail:kato@is.naist.jp

1.1 Subjective Measurements in HAR Evaluations

Researchers have used standardized questionnaires such as the System Usability Scale (SUS) [Lewis and Sauro 2009] and the NASA Task Load Index (TLX) [Hart 2006] to conduct subjective measurements. These questionnaires have been shown to be valid and reliable measures, and enable three types of comparisons:

(a) Between iterations – comparisons between past version to current version of the same application in iterative prototyping;

(b) Between features – comparisons between different features of an interface with the goal of identifying improvements to prioritize;

(c) Benchmarking – comparing the HAR application to the state-of-the-art, or simply some other implementation by another group.

Although widely used for these three comparisons, the SUS and NASA-TLX do not represent specific perceptual and ergonomic issues common to HAR. As such, researchers complement these tools with their own questionnaires. However, these questionnaires are not always tested for validity and reliability. Moreover, the questions are too specific to the features of their HAR application. Thus, these questionnaires cannot be used for comparison (b) between features nor comparison (c) benchmarking.

1.2 Contribution

In response to the lack of evaluation tools specific for HAR, the main contribution of this paper is the creation of the HAR Usability Scale (HARUS). Researchers and professionals involved in developing HAR applications can take advantage of our usability scale for comparative evaluations of their systems. HARUS is general enough for most HAR applications while considering specific perceptual and ergonomics issues common in HAR. We based this on issues reported by actual users from our systematic literature review. More importantly, the results of our experiments show that HARUS is valid because it measures usability and that HARUS is reliable, that is, the questions are consistent.

2 Evaluation of HAR

The main points of evaluation in mobile AR are in its unique perceptual and ergonomic issues [Swan and Gabbard 2003]. At the time, mobile AR involves carrying computers with backpacks while wearing head-mounted displays and other peripherals. Currently, mobile AR can also be implemented in single handheld device. This new enabling technology led to new perceptual and ergonomic issues that need to be evaluated in user-based studies [Kurkovsky et al. 2012].

2.1 Perceptual Issues

Several researchers have studied perceptual issues in AR with respect to enabling devices: head-mounted display (HMD), handheld device, and projector-camera system. Kruijff et al. considered the human visual processing and interpretation pipeline, and carefully summarized these perceptual issues [2010]. They associated these perceptual issues to implementation issues found in: the real environment, capturing, augmentation, display and individual user differences. Moreover, they described several issues and disadvantages arising from the form of handheld devices.

Handheld devices include cellular phones, smart phones, tablet computers, ultra-mobile computers, etc. Currently, many handheld devices have powerful processors, large LCD screens, and internal cameras. These features allow researchers to implement AR in one compact device. Although handheld devices are useful for many applications, Kruijff et al. listed the following disadvantages: less visibility of the LCD screen, lower fidelity, different disparity planes, higher latency, and smaller screen size in contrast with HMD and projector-camera systems [2010]. Through our systematic literature review, we support these insights with perceptual issues reported by actual users when testing HAR applications.

2.2 Ergonomic Issues

Aside from perceptual issues unique to HAR, we must also consider ergonomic issues specific to the behavior of people using HAR. Among the several interactions afforded by HAR, the most common use is the magnifying glass metaphor [Rekimoto 1997]. In this interaction style, the users hold the handheld device in front of them. The screen faces them, and the camera points to a scene. This kind of interaction is very different from conventional uses of handheld devices. As such, previous tools used for mobile devices such as the Mobile Phone Usability Questionnaire (MPUQ) [Ryu and Smith-Jackson 2006] do not evaluate such interaction. Although there are some overlaps between MPUQ and HARUS, especially in questions related to cognitive load and control, standard questionnaires for mobile phones do not give emphasis on fatigue associated with the unique gestures in using HAR.

Veas and Kruijff evaluated several handheld platforms to understand and address these ergonomic issues [2010]. They describe HAR ergonomics issues to be an interplay of issues in pose, grip, controller allocation, weight and size. The goal of their design is for the users to hold a particular pose while gripping the handheld device. Aside from viewing interactions, they also considered input interactions such as having additional controllers. As expected, the size and weight of the device and the whole system are important considerations in HAR. Through our systematic literature review, we support these insights with ergonomic issues reported by actual users when testing HAR applications.

2.3 Design Goals

Given these two types of issues, it follows that the goal of design for HAR is to have no perceptual and ergonomics issues. In this paper, we refer to these qualities as *comprehensible* and *manipulable*, respectively. In other words, a perfect HAR application would score 100% on measures of comprehensibility and manipulability. In this paper, we approximate HAR usability to be equivalent to a linear combination of comprehensibility and manipulability. We assume usability to be the average of these factors, and we then provided evidences that these are sound estimations.

3 Approach

To create the HAR Usability Scale (HARUS), we followed a five-step method for developing and testing questionnaires or instruments [Radhakrishna 2007]. The five steps are background, questionnaire conceptualization, format and data analysis, establishing validity and establishing reliability.

1. Studying the background – We conducted a systematic literature review to explore the common problems experienced by users when using HAR applications.

2. Conceptualizing the questionnaire – We defined two factors that we want to measure with our questionnaire: *comprehensibility* and *manipulability*. Comprehensibility is the ease of understanding the information presented by the HAR system. On the other hand, manipulability is the ease of handling the

HAR system as the user performs the task. Comprehensibility and manipulability correspond to the perceptual and ergonomic issues in HAR, respectively. Thus, we assume that the usability of a HAR system is approximated by comprehensibility and manipulability factors. Our questionnaire is patterned from the SUS [Lewis and Sauro 2009], and follows the design rules prescribed by Fowler and Cosenza [2008].

3. Choosing the format and data analysis – We designed the questionnaires to be answerable using Likert scales, similar to the SUS. In other words, we asked users to indicate how much they agree or disagree to the statement presented to them by rating a scale from 1 to 7. Only 1 and 7 are labeled, with 1 labeled as "Strongly Disagree" and 7 labeled as "Strongly Disagree". We used a 7-point Likert scale because the audience of our experiments are sophisticated enough to distinguish subtle differences in these scales as recommended by Krosnick and Presser [2010]. We ordered the questions such that we alternate between positively stated and negatively stated questions.

4. Establishing validity - We validated our HARUS by showing concurrent validity, a kind of criterion-oriented validation procedure [Cronbach and Meehl 1955]. Concurrent validity is demonstrated when a test correlates well with objective measurements (time on task, etc.) or subjective measurements (SUS, etc.) that has been previously validated. As such, validity is a matter of degree, not all or nothing [Messick 1990]. Intuitively, we know that HARUS should correlate with other subjective measurements of usability because of its design. However, it is interesting to know the strength of the correlation in actual experiments.

5. Establishing reliability - We measured the reliability or the precision of HARUS by computing the Cronbach's alpha – a measure of internal consistency of a questionnaire [Krosnick and Presser 2010].

4 Systematic Literature Review

We used the search string *handheld AND "augmented reality" AND evaluation* to search the ACM Digital Library for relevant research papers. This search resulted in 959 papers which we narrowed to 10 articles (column 2 of Table 1) by applying the following inclusion criteria:

1. Must discuss a HAR application
2. Must conduct a user-based evaluation
3. Must be the latest article on that HAR application

We read the papers with focus on listing issues raised by users, and issues observed by experimenters or expert reviewers as actual users use the system. We listed these issues encountered by users in Table 1 (first column). We then conceptualized measures (Table 2) that are statements based on these issues.

5 HAR Usability Scale

We score HARUS similar to the SUS [Bangor et al. 2008]. We designed it to have a two-factor structure representing comprehensibility and manipulability. For each factor, multiple questions are asked to help the users evaluate various aspects contributing to their experience with the HAR. The HARUS is intended to measure the usability of a HAR given a target user group and a confined task.

5.1 Scoring Method

The HARUS is composed of 16 statements (Table 2) that roughly correspond to commonly encountered problems in HAR applications. Users were then asked to rate their agreement by using a 7-point Likert scale. To compute the HARUS score, we apply a similar method for computing the SUS score [Bangor et al. 2008]:

1. For the positively-stated items, subtract one from the user response. For the negatively-stated items, subtract the user response from seven.
2. Add all these converted responses.
3. Divide the sum by 0.96 to have a score ranging from 0 to 100.

5.2 Factor Structure

The HARUS has a two-factor structure. Statements 1 to 8 are measures of manipulability, whereas statements 9 to 16 are measures of comprehensibility. However, these statements are not an exhaustive operationism of manipulability or comprehensibility. Rather, they are measures belonging to an extensible set of indicators for these two constructs [Messick 1995]. Similarly, we do not claim that these 16 questions, and two constructs are the strict operationism of usability in HAR. They are measures belonging to an extensible set of indicators for usability. However, we showed evidence in the succeeding three experiments that this set of measures is a good approximation of usability in HAR.

5.3 Multiple Measures

The SUS [Lewis and Sauro 2009] is composed of 10 statements that breaks the question "Is this system easy to use?" into several aspects of the system. Similarly, the concept of HARUS is to break down the questions "Is this system easy to handle?" and "Is the information presented easy to understand?" so that the users find it easier to give their feedback. When users are asked general questions like "Is this system easy to use?", they would not know how to weigh various aspects of the system to come up with a single rating. They can give better feedback if they can rate smaller, more specific aspects. These ratings can then be accumulated to gauge their answer to the bigger, general questions.

5.4 Generality

There are many areas of HAR applications, including advertising, navigation, work support, scientific visualization, etc. Some may argue that the main factors affecting usability will vary according to the application area. Some may say that the purpose of the HAR application is different, thus the requirements are different. We offer two arguments why HARUS can be used to all application areas.

First, HARUS is not intended to give an overall evaluation of a HAR application. Rather, it evaluates the suitability of HAR application to target users and tasks. Usability evaluations are always with respect to the user and the task [Nielsen 1994]. Some HAR application areas would have more tech-savvy target users. Some would have tasks that require more dexterity. Depending on the application areas, researchers decide on their target users and tasks when conducting usability evaluation. However, if the goal is to measure how well users can use the HAR application for a task, then researchers can still use HARUS. For example, a sophisticated HAR application for work support might have a lower HARUS score than a crude HAR application for advertising because the tasks in work support are more difficult. This is fair because it is possible to create a crude application that addresses the needs of a user for a specific

Table 1: *User Issues in HAR Applications and Corresponding HARUS Statements*

Issue	References	Statements (Table 2)
The tracking is unstable due to the ambient light, bad sensor fusion, or mishandling of the user.	[Lee et al. 2012], [Dünser et al. 2012], [White and Feiner 2009], [Mulloni et al. 2011a], [Olsson and Salo 2011]	(15), (16)
The virtual objects are not well-registered.	[Lee et al. 2012], [Dünser et al. 2012], [White and Feiner 2009], [Mulloni et al. 2011a], [Mulloni et al. 2011b]	(13)
The application is lagging or has intolerable latency.	[Lee et al. 2012]	(12)
The content was excessive and has poor quality.	[White and Feiner 2009], [Mulloni et al. 2011b], [Olsson and Salo 2011]	(10), (13)
The display induces much cognitive load.	[Dünser et al. 2012], [Olsson and Salo 2011]	(9), (10)
The download time of the content is too slow.	[Olsson and Salo 2011]	(12)
The screen is not legible due to outdoor ambient light.	[Dünser et al. 2012], [Schall et al. 2009], [Veas et al. 2013], [Dey and Sandor 2014]	(11), (14)
The screen is not legible due to reflection or glare.	[Lee et al. 2012], [Veas et al. 2013]	(11), (14)
Depth is not understood or underestimated.	[Schall et al. 2009], [Dey and Sandor 2014]	(13)
The application causes fatigue after extended use.	[Veas and Kruijff 2008], [Schall et al. 2009]	(5)
The device is too bulky or too heavy.	[Veas and Kruijff 2008], [Veas et al. 2013]	(1), (2)
Hand interactions are difficult to perform.	[Veas and Kruijff 2008], [Veas et al. 2013]	(3), (7),(8)
The application is not responsive or provides no feedback.	[Lee et al. 2012], [Olsson and Salo 2011]	(6), (8), (12)
The keypad is too small.	[Veas et al. 2013]	(4)

Table 2: *The HAR Usability Scale*

	Manipulability Measures:	Relevance to HAR:
1	I think that interacting with this application requires a lot of body muscle effort.	HAR is often used while moving around the real environment.
2	I felt that using the application was comfortable for my arms and hands.	HAR strains the hands and arms the most.
3	I found the device difficult to hold while operating the application.	HAR has grip and pose issues.
4	I found it easy to input information through the application.	HAR introduces novel interaction metaphors.
5	I felt that my arm or hand became tired after using the application.	HAR strains the hands and arms the most.
6	I think the application is easy to control.	HAR introduces novel interaction metaphors.
7	I felt that I was losing grip and dropping the device at some point.	HAR has grip and pose issues.
8	I think the operation of this application is simple and uncomplicated.	HAR introduces novel interaction metaphors.
	Comprehensibility Measures:	Relevance to HAR:
9	I think that interacting with this application requires a lot of mental effort.	HAR is susceptible to presenting too much information on a small screen.
10	I thought the amount of information displayed on screen was appropriate.	HAR introduces novel visualization metaphors.
11	I thought that the information displayed on screen was difficult to read.	HAR has legibility issues due to ambient light, glare, etc.
12	I felt that the information display was responding fast enough.	HAR has latency issues due to the limited processing power and network connection.
13	I thought that the information displayed on screen was confusing.	HAR introduces novel visualization metaphors.
14	I thought the words and symbols on screen were easy to read.	HAR has legibility issues due to ambient light, glare, etc.
15	I felt that the display was flickering too much.	HAR is susceptible to tracking and registration errors due to many factors, such as dynamics of lighting.
16	I thought that the information displayed on screen was consistent.	HAR is susceptible to tracking and registration errors due to many factors, such as dynamics of lighting.

task, and it is also possible to create a sophisticated application that does not. The evaluation is relative to the user groups and the tasks.

Second, we applied the best effort because we considered issues in as much application areas as possible. We based HARUS only on known issues because we cannot predict future issues that will arise in new application areas. These known issues will still be the problem in HAR applications in the next coming years. Furthermore, we applied multiple experiments with multiple benchmarks including both objective and subjective measures of usability. The experiments are both practical and general.

6 Experiment 1: Annotating Text

In this experiment, users evaluated an application for annotating text on real objects found in the environment.

Specific HAR applications aim to create AR content in situ. In the work of Langlotz et al. [2012], the users create virtual content directly onto the environment by using only a smartphone. The usability of such an authoring system can be evaluated using the SUS [Lewis and Sauro 2009], although some information considering perceptual and ergonomics issues are lost. We can also use time on task to evaluate this system because people who will find the application difficult to use would tend to finish the task with more time.

This experiment tests the ability of HARUS in evaluating a simple HAR content authoring task. We evaluated the concurrent validity of HARUS by checking its correlation with SUS, a previously validated subjective measure. Furthermore, we benchmarked against time on task, an objective measure. Our hypotheses are as follows:

H1. HARUS and SUS have a positive relationship.

H2. HARUS and time on task have a negative relationship.

6.1 Experimental Platform

We implemented a simple HAR authoring tool for annotating text on real objects (Figure 2, right). We used the PointCloud SDK[1] to detect some natural feature points in the environment. To register feature points, the user must move the handheld device from side-to-side (Figure 2, left). Once the system detects enough feature points, the user can add a text label on to the scene.

The application runs on iPad 2 tablets (A5 processor, 512MB DDR2 RAM, 32GB, 601 grams). We used the back camera (640x480 pixels, 30 fps) for sensing, and 9.7 inch LED-display (1024x768 at 132 ppi) for display. The interface was built using standard interface elements of iOS 6 such as labels, textfields, keyboard, etc. as shown in Figure 2 (middle).

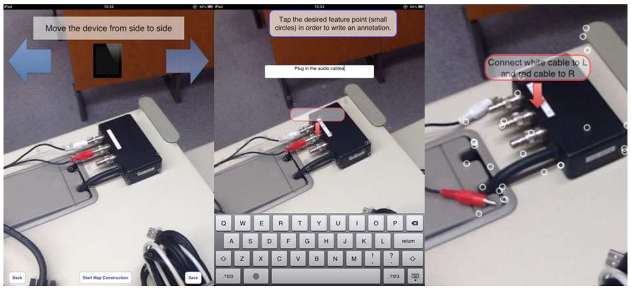
Move the device from side to side
Tap the desired feature point (small circles) in order to write an annotation.
Plug in the audio cables
Connect white cable to L and red cable to R
Back
Start Map Construction
Save

Figure 2: *Simple HAR Authoring Tool for Annotating Text*

6.2 Design and Procedure

Eighteen voluntary participants with ages ranging from 22 to 41 years (M=27, SD=4.0) participated in this experiment. First, the experimenters demonstrated how to use the authoring tool. The participants were then asked to annotate English translations on a rice cooker, and annotate trivia on a paper bill (Figure 3). No time limit was given to do the tasks, and the participants were free to give up. 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 HARUS questionnaires. Nine answered the SUS first, whereas nine answered HARUS first. We took note of the time on task, and we calculated the HARUS and SUS scores as described in Section 5.1.

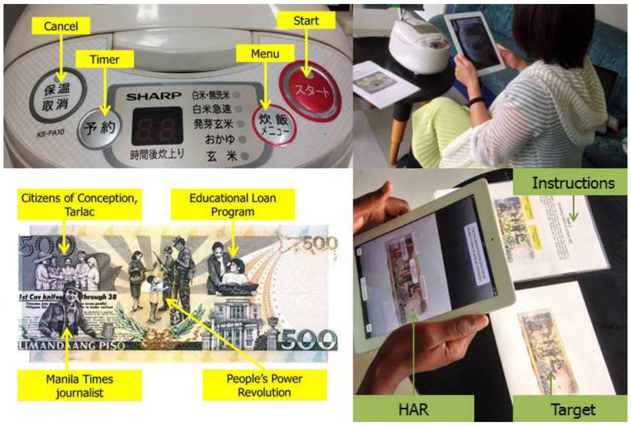

Figure 3: *Authoring Tasks*

6.3 Results

Fifteen participants finished the task with an average time of 8.1 minutes (SD=2.5). The 18 participants gave the HAR authoring tool an average SUS score of 62 (SD=22) and an average HARUS score of 65 (SD=16). These scores are below the acceptable SUS score of 70 and above [Bangor et al. 2008].

The HARUS score has a very strong positive relationship with the SUS score and a strong negative relationship with time on task (Table 3). Both findings are significant, thereby supporting hypotheses 1 and 2. These findings are indicative of the concurrent validity of HARUS.

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*	-0.59**	1.00

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

[1] http://developer.pointcloud.io/sdk

7 Experiment 2: Status Reporting

In this experiment, users evaluated an application for viewing virtual notes on real devices for writing a report.

HAR applications commonly require users to read virtual information associated with real environments. This information could be an advertisement, scientific data, historical information, etc. Aside from the SUS, this kind of commercial application can be evaluated based on the "Affective Aspects and Media Properties" (AAMP) construct of the MPUQ [Ryu and Smith-Jackson 2006]; such that, an easy to use product would elicit positive emotional responses. Of the 14 statements measuring AAMP, we chose 8 that were relevant to our task.

Furthermore, in work-related tasks, a useful HAR should lead to better work output. This experiment also checked the relationship of the HARUS with the verbosity of the status report. We assume that writing more words on the report means that the report is more comprehensive and is thus of better quality. Our hypotheses are:

H3. HARUS and SUS have a positive relationship.

H4. HARUS and AAMP have a positive relationship.

H5. HARUS and report verbosity have a positive relationship.

7.1 Experimental Platform

The HAR application enables users to view text annotations on real objects (Figure 4, left and middle). The application runs on iPad mini tablets (A7 processor, 512MB DDR2 RAM, 16GB, 308 grams). We used the back camera (640x480 pixels, 30 fps) for sensing, and a 7.9 inch LED-display (1024-by-768 at 163 ppi). We used the PointCloud SDK for tracking, and the standard user interface elements of iOS 7 for the display.

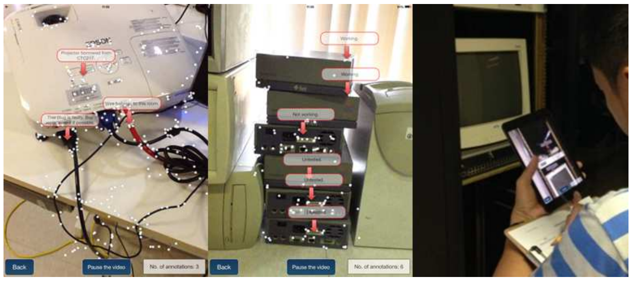

Figure 4: *HAR for Viewing Annotations on Equipment*

7.2 Design and Procedure

Twenty voluntary participants with ages ranging from 19 to 46 years (M=28, SD=8.1) participated in this experiment. Before performing the task, we explained AR and its enabling technologies to the participants by using videos and slides. We then demonstrated how to use the HAR application for viewing text annotations. The participants assumed the role of a newly-hired maintenance staff. Their first job is to report on the status of equipment by viewing the annotations made by the previous maintenance staff. They then filled the report form that has three columns: device, description of 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. We gave them a time limit of 15 minutes to finish the task. We finally asked them to answer three questionnaires: HARUS, SUS and AAMP. We computed an AAMP score similar to the method for SUS.

7.3 Results

This kind of work-support task is not limited to those that use head-mounted display. Several researchers apply HAR because it is less intimidating and easier to share, thereby facilitating collaboration with co-workers [Schall et al. 2009]. This task is suitable for HAR because writing the report requires both information displayed by HAR, and information gathered from the real environment such as the description of the device. The natural interaction pattern we observed 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 writing the report.

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) which is an acceptable SUS score. 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 AAMP scores (Table 4). Both results are significant, thereby supporting hypotheses 3 and 4. These results are indicative of the concurrent validity of HARUS. We did not find any significant relationship between HARUS and verbosity probably because low word count could mean both lacking in information (bad quality) or simply concise (good quality). For our future work, we plan to use more sophisticated methods of measuring the quality of written reports.

Table 4: *Correlations (r) of HARUS, SUS, AAMP and Verbosity*

	1	2	3	4
1. HARUS	1.00			
2. SUS	0.79***	1.00		
3. AAMP	0.75***	0.82***	1.00	
4. Verbosity	0.12	0.23	0.43*	1.00

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

8 Experiment 3: Learning Words

In this experiment, users evaluated a HAR application for learning Filipino words from a real environment.

Researchers have suggested to use HAR in mobile learning [Kamarainen et al. 2013] and other AR learning experiences. HAR transforms the real environment into a learning experience by adding virtual content onto it. Moreover, HAR is used to direct the attention of the users to otherwise unnoticeable details in the environment. It can support collaborative learning, embodied cognition and contextual visualization [Santos et al. 2014].

This experiment tests the ability of HARUS in evaluating a HAR application in an educational scenario. Aside from comparing with the SUS, we demonstrated concurrent validity by checking the correlation of HARUS with the Instructional Material Motivation Survey (IMMS) [Huang et al. 2006]. This questionnaire was designed to measure how motivating an instructional material is. As Pilke described, bad usability interferes with flow, which may lead to frustrated users [2004]. On the other hand, good usability engages the user and keeps them motivated to perform the task well.

Furthermore, we compared HARUS to total study time, which is an objective measure of motivation. In other words, users who find the learning material motivating would tend to study more. Our hypotheses are as follows:

H6. HARUS and SUS have a positive relationship.

H7. HARUS and IMMS have a positive relationship.

H8. HARUS and total study time have a positive relationship.

8.1 Experimental Platform

We implemented a simple word acquisition application on iPad 2 tablets with similar specification as in Experiment 1. We used ARToolkit[2] to measure the pose of the camera with respect to the target object. Thus, we attached fiducial markers to each of the objects that we want to track. We then rendered the virtual content by using OpenGL ES 2.0[3]. Figure 5 shows the 3-D registered sprite sheet animation. The HAR application can display some text descriptions and play the proper pronunciations of the word.

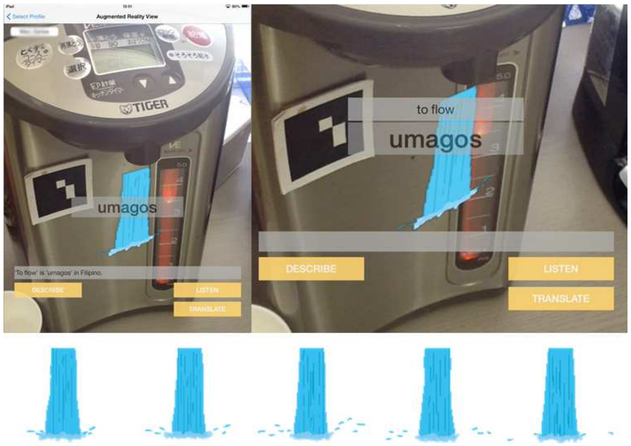
'To flow' is 'umagos' in Filipino.

Figure 5: *HAR Application for Word Acquisition in a Real Environment. The sprite sheet animation illustrate the word "umagos" which is the Filipino word for "to flow."*

8.2 Learning Environment

We set up our HAR for a 5x3 meter refreshment area of a laboratory. Members of the laboratory use this area for eating snacks, making hot drinks, reading comic books, chatting, etc. We tagged 30 objects found in the environment by attaching fiducial markers as shown in Figure 6. 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. In other words, we avoided languages that are close to those that our participants are already familiar with. Objects used to teach Filipino nouns are annotated with the word itself as virtual labels. Those teaching Filipino verbs are annotated with sprite sheets (Figure 5) that demonstrate the corresponding action.

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). The buttons can be pushed as much as the user wants.

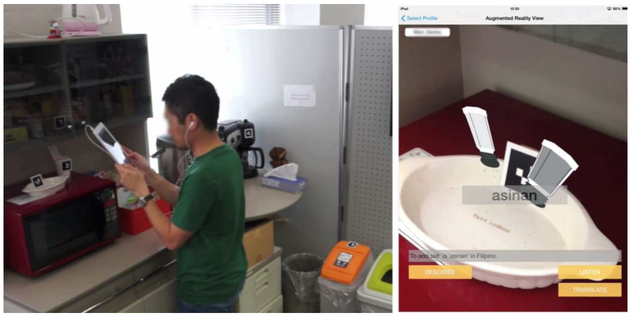
'To add salt' is 'asinan' in Filipino.

Figure 6: *Learning Words in a Real Environment*

8.3 Design and Procedure

Eighteen voluntary participants with ages ranging from 23 to 32 years (M=26, SD=2.6) participated in this experiment. All of them are familiar with AR. Ten of the participants have used some mobile phone application for learning second languages. None of them are familiar with the Filipino language. 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 participant has a user account which we monitored by logging activities and usage of each account. After the last day of studying, we asked the participants to answer three questionnaires: HARUS, SUS and IMMS. We computed an IMMS score similar to SUS.

8.4 Results

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) which is an acceptable usability score. However, they rated the application 61 (SD=15) on the HARUS. This difference is the largest we observed between the SUS and HARUS scores in the three experiments. 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. In other words, participants who gave higher HARUS scores tend to find the interface more motivating. They also tend to study more with the interface. All of these correlations are significant and support hypotheses 6 to 8. These results are indicative of the concurrent validity of HARUS.

Table 5: *Correlations (r) of HARUS, SUS, IMMS and Study Time*

	1	2	3	4
1. HARUS	1.00			
2. SUS	0.68^{***}	1.00		
3. IMMS	0.61^{**}	0.55^{**}	1.00	
4. Study Time	0.42^{*}	0.49^{*}	0.13	1.00

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

[2] http://www.hitl.washington.edu/artoolkit/
[3] http://www.opengl.org/

9 Reliability of HARUS

After showing evidence of the validity of HARUS, the last step in developing an evaluation tool is to measure its reliability. We computed the Cronbach's alpha to assess the internal consistency of our questionnaire based on the responses in the experiments in Sections 6 to 8. All of our alphas are between 0.7 to 0.9. Thus HARUS has good internal consistency as shown in Table 6.

Table 6: *Cronbach's Alpha (α) in Three Experiments*

Experiment	α	Interpretation
Annotating Text	0.83	Good
Status Reporting	0.83	Good
Learning Words	0.87	Good

10 Manipulability and Comprehensibility

Through the three experiments, we showed the validity and reliability of HARUS. We demonstrated concurrent validity by providing evidences supporting seven out of our eight hypotheses. In all three experiments, the participants were able to answer the questions consistently as measured by the Cronbach's alpha. Aside from these main findings, we explored some interesting relationships between the factors of HARUS and other variables in our experiment.

10.1 Two-factor Structure of HARUS

The HARUS can be decomposed into two scores: the manipulability score based on questions 1 to 8, and the comprehensibility score based on questions 9 to 16 (Table 2), which was computed similarly as the HARUS score (Section 5). In previous user studies of HAR applications, perceptual and ergonomic issues are described to be interrelated in some user issues mainly because the manner of handling the device affects the quality of visualization. Moreover, instability in tracking makes the tasks longer. As such, participants report fatigue especially in the arms and hands.

In our experiments, we observed manipulability and comprehensibility to be interrelated but moderate enough to be used as different scales. Table 7 summarizes the correlations of these two factors of HAR usability. We only observed a strong positive relationship in the learning words scenario. For both the annotating text and status-reporting tasks, the correlations were moderate and not significant. As such, our guess is that manipulability and comprehensibility in HAR depend on the users and on the tasks. Therefore, these two factors must be observed independently from each other. Our current HAR usability scale is also suitable for this observation.

Table 7: *Correlations (r) of Manipulability and Comprehensibility in Three Experiments*

Experiment	Pearson's r	Interpretation
Annotating Text	0.40	moderate
Status Reporting	0.34	moderate
Learning Words	0.60**	strong

** significant at 0.01 level

10.2 Relationships of Manipulability and Comprehensibility with Experiment Variables

Scoring the manipulability and comprehensibility factors of the HARUS reveals additional insights from the three experiments. In all three experiments, comprehensibility has a stronger positive relationship with SUS compared to manipulability. The difference was small for the authoring scenario (Table 8). However, the difference was very pronounced for the viewing text (Table 9) and learning words scenario (Table 10) most probably because of the nature of the task. In the authoring text annotations scenario, manipulability was very important to the whole usability of the interface because it required the users to do difficult hand movements such as moving the application from side-to-side to register enough feature points, positioning labels, and typing some text. On other hand, these input interactions are not considered in the status reporting and learning words tasks. The focus of these tasks was to understand the information presented to the user.

Table 8: *Correlations (r) of HARUS Factors, SUS and Time on Task in Annotating Text Scenario*

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

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

Table 9: *Correlations (r) of HARUS Factors, SUS, AAMP and Verbosity in Status Reporting Scenario*

	1	2	3	4	5
1. Manipulability	1.00				
2. Comprehensibility	0.34	1.00			
3. SUS	0.58**	0.70***	1.00		
4. AAMP	0.54*	0.68***	0.82***	1.00	
5. Verbosity	0.41*	-0.19	23	0.43*	1.00

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

Table 10: *Correlations (r) of HARUS Factors, SUS, IMMS and Study Time in Learning Words Scenario*

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

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

In the status-reporting task, we did not find strong positive relationship between the HARUS score and the verbosity of the reports. However, a strong positive relationship exists between the manipulability score and verbosity. In other words, people who found the HAR easy to handle tend to write more words on their report. We find this to be logical, and we believe that there are trade-offs in user performance for activities that use the hands (e.g. handling the HAR and hand-writing a report). We plan to investigate this more in our next experiments.

Aside from the difference in SUS scores, we also observed differences in the strength of correlations in the two factors for the learning words task. Comprehensibility has a strong positive relationship with study time, but not manipulability. In other words, those who found the HAR leaning material easy to understand tend to study more. Intuitively, one would guess that the usability of a HAR learning material would be affected more by the comprehensibility of the information presented rather than manipulability.

10.3 Reliability of Manipulability and Comprehensibility Scales

Given that the manipulability and comprehensibility statements can be used as separate scales, we evaluated the internal consistency of the responses in three experiments. Although the Cronbach's alpha are slightly lower than those of HARUS, both manipulability and comprehensibility have good internal consistency as shown in Table 11. Thus, these two HARUS factors can be used as separate scales in cases wherein researchers are only interested to measure these factors.

Table 11: *Cronbach's Alpha of Manipulability and Comprehensibility in Three Experiments*

Experiment		α	Interpretation
Annotating Text	Manipulability	0.71	Good
	Comprehensibility	0.74	Good
Status Reporting	Manipulability	0.81	Good
	Comprehensibility	0.80	Good
Learning Words	Manipulability	0.83	Good
	Comprehensibility	0.79	Good

11 Summary of Findings

We conducted three experiments to measure the validity and reliability of HARUS. Furthermore, we explored the insights found in its decomposition into separate comprehensibility and manipulability scales. We enumerate the findings as follows:

1. In all three experiments, HARUS and SUS have a significantly very strong positive relationship.
2. In experiment 1, the HARUS score increases as the time taken to finish the task decreases.
3. In experiment 3, the students who gave higher HARUS scores tend to study for longer periods of time.
4. In experiments 2 and 3, the given HARUS score increases with self-report measures of positive emotions and motivation, respectively.
5. In experiment 3, we observed a large margin between the HARUS and SUS scores. We gained an acceptable SUS score, however the HARUS score is low. We believe that this is because the SUS does not capture the problems unique to HAR. However, this should be further investigated through user studies.
6. In all experiments, the HARUS demonstrated good internal consistency. Similarly, the separate manipulability and comprehensibility scales also have good internal consistency in all our experiments.
7. The manipulability and comprehensibility scales have varying degrees of relationship strength with the SUS, time on task, study time, positive emotions and motivation. These separate scales provide more insight when analyzing HAR.
8. HARUS can be decomposed into separate manipulability and comprehensibility scales. These constructs should be analysed separately because they ony correlate moderately in some cases. In other words, it is possible for a HAR to suffer more from manipulability issues than comprehensibility issues, and vice versa, but not at the same time.

12 Conclusion

HARUS and its factors - the manipulability scale and the comprehensibility scale - are tools for evaluating HAR applications with users as they perform specific tasks. Such evaluation tools are important to researchers and professionals in measuring the usability of their HAR applications. Aggregating usability as a single usability score allows them to compare between iterations of the same application, to prioritize among several features of an application, and to benchmark against previously evaluated implementations of HAR.

HAR is a novel interface that has high potential for becoming a mainstream technology. It is useful for delivering various content in many fields of application. The improvements in the enabling technology and development of specific applications must be accompanied by developing new ways to assess such interfaces to ensure commercial success.

In this paper, we presented our experiences in creating the HAR Usability Scale. Most importantly, we provided evidences of the soundness of this technique. We find this tool useful in evaluating our interfaces. However, this tool must also be used by other researchers so that we can understand more the extents of its applicability as well as its limitations.

Acknowledgements

We would like to thank Arno in Wolde Lübke for helping implement some of the prototypes, Raphael C. Santos for drawing our assets, and Prof. Ma. Mercedes T. Rodrigo for sharing her insights to our learning scenario and experiments.

References

BANGOR, A., KORTUM, P. T., AND MILLER, J. T. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction 24*, 6, 574–594.

CRONBACH, L. J., AND MEEHL, P. E. 1955. Construct validity in psychological tests. *Psychological bulletin 52*, 4, 281.

DEY, A., AND SANDOR, C. 2014. Lessons learned: Evaluating visualizations for occluded objects in handheld augmented reality. *International Journal of Human-Computer Studies 72*, 1011, 704 – 716.

DÜNSER, A., GRASSET, R., AND BILLINGHURST, M. 2008. *A survey of evaluation techniques used in augmented reality studies*. Human Interface Technology Laboratory New Zealand.

DÜNSER, A., BILLINGHURST, M., WEN, J., LEHTINEN, V., AND NURMINEN, A. 2012. Exploring the use of handheld ar for outdoor navigation. *Computers & Graphics 36*, 8, 1084 – 1095. Graphics Interaction Virtual Environments and Applications 2012.

FOWLER, F. J., AND COSENZA, C. 2008. *International Handbook of Survey Methodology*. Taylor & Francis, ch. Writing Effective Questions, 136–159.

GABBARD, J., AND SWAN, J. 2008. Usability engineering for augmented reality: Employing user-based studies to inform design. *Visualization and Computer Graphics, IEEE Transactions on 14*, 3 (May), 513–525.

GERVAUTZ, M., AND SCHMALSTIEG, D. 2012. Anywhere interfaces using handheld augmented reality. *Computer 45*, 7 (July), 26–31.

HART, S. G. 2006. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, Sage Publications, 904–908.

HIX, D., GABBARD, J., SWAN, J.E., I., LIVINGSTON, M., HOLLERER, T., JULIER, S., BAILLOT, Y., AND BROWN, D. 2004. A cost-effective usability evaluation progression for novel interactive systems. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, 10 pp.–.

HUANG, W., HUANG, W., DIEFES-DUX, H., AND IMBRIE, P. K. 2006. 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 37*, 2, 243–259.

KAMARAINEN, A. M., METCALF, S., GROTZER, T., BROWNE, A., MAZZUCA, D., TUTWILER, M. S., AND DEDE, C. 2013. Ecomobile: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education 68*, 0, 545 – 556.

KROSNICK, J. A., AND PRESSER, S. 2010. *Handbook of Survey Research*. Emerald Group Publishing Limited, ch. Question and Questionnaire Design, 263–313.

KRUIJFF, E., SWAN, J., AND FEINER, S. 2010. Perceptual issues in augmented reality revisited. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, 3–12.

KURKOVSKY, S., KOSHY, R., NOVAK, V., AND SZUL, P. 2012. Current issues in handheld augmented reality. In *Communications and Information Technology (ICCIT), 2012 International Conference on*, 68–72.

LANGLOTZ, T., MOOSLECHNER, S., ZOLLMANN, S., DEGENDORFER, C., REITMAYR, G., AND SCHMALSTIEG, D. 2012. Sketching up the world: in situ authoring for mobile augmented reality. *Personal and Ubiquitous Computing 16*, 6, 623–630.

LEE, G. A., DUNSER, A., KIM, S., AND BILLINGHURST, M. 2012. Cityviewar: A mobile outdoor ar application for city visualization. In *Mixed and Augmented Reality (ISMAR-AMH), 2012 IEEE International Symposium on*, 57–64.

LEWIS, J. R., AND SAURO, J. 2009. The factor structure of the system usability scale. In *Human Centered Design*. Springer, 94–103.

MESSICK, S. 1990. Validity of test interpretation and use. Tech. rep., Educational Testing Service.

MESSICK, S. 1995. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist 50*, 9, 741.

MULLONI, A., SEICHTER, H., AND SCHMALSTIEG, D. 2011. User experiences with augmented reality aided navigation on phones. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, 229–230.

MULLONI, A., SEICHTER, H., AND SCHMALSTIEG, D. 2011. Handheld augmented reality indoor navigation with activity-based instructions. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, ACM, New York, NY, USA, MobileHCI '11, 211–220.

NIELSEN, J. 1994. *Usability Engineering*. Elsevier.

OLSSON, T., AND SALO, M. 2011. Online user survey on current mobile augmented reality applications. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, 75–84.

PILKE, E. M. 2004. Flow experiences in information technology use. *International journal of human-computer studies 61*, 3, 347–357.

RADHAKRISHNA, R. B. 2007. Tips for developing and testing questionnaires/instruments. *Journal of Extension 45*, 1, 1–4.

REKIMOTO, J. 1997. A magnifying glass approach to augmented reality systems. *Presence 6*, 4, 399–412.

RYU, Y. S., AND SMITH-JACKSON, T. L. 2006. Reliability and validity of the mobile phone usability questionnaire (mpuq). *Journal of Usability Studies 2*, 1, 39–53.

SANTOS, M., CHEN, A., TAKETOMI, T., YAMAMOTO, G., MIYAZAKI, J., AND KATO, H. 2014. Augmented reality learning experiences: Survey of prototype design and evaluation. *Learning Technologies, IEEE Transactions on 7*, 1 (Jan), 38–56.

SCHALL, G., MENDEZ, E., KRUIJFF, E., VEAS, E., JUNGHANNS, S., REITINGER, B., AND SCHMALSTIEG, D. 2009. Handheld augmented reality for underground infrastructure visualization. *Personal and Ubiquitous Computing 13*, 4, 281–291.

SWAN, J., AND GABBARD, J. L., 2003. Perceptual and ergonomic issues in mobile augmented reality for urban operations.

VEAS, E., AND KRUIJFF, E. 2008. Vesp'r: design and evaluation of a handheld ar device. In *Mixed and Augmented Reality, 2008. ISMAR 2008. 7th IEEE/ACM International Symposium on*, 43–52.

VEAS, E. E., AND KRUIJFF, E. 2010. Handheld devices for mobile augmented reality. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, ACM, 3.

VEAS, E., GRASSET, R., FERENCIK, I., GRNEWALD, T., AND SCHMALSTIEG, D. 2013. Mobile augmented reality for environmental monitoring. *Personal and Ubiquitous Computing 17*, 7, 1515–1531.

WHITE, S., AND FEINER, S. 2009. Sitelens: Situated visualization techniques for urban site visits. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '09, 1117–1120.