Adaptation of the System Usability Scale for User Testing with Children

Cynthia Putnam
Melisa Puthenmadom
Marjorie Ann Cuerdo
Wanshu Wang
Nathaniel Paul
College of Computing and Digital Media
DePaul University, Chicago, IL, USA
cputnam@depaul.edu
puthenm2@gmail.com
mcuerdo@mail.depaul.edu
wanshu.wang@gmail.com
nathanpaul714@gmail.com

CHI'20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA

ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3382840

Abstract
In this paper, we describe a pilot study in which we adapted and tested the System Usability Scales (SUS) for children between ages of 7-11. We began the study with interviews with four elementary school teachers in which we asked their help with modifying the SUS usability statements for children. We then tested those questionnaire statements with 30 children after they completed puzzles in mobile apps; we assessed the statements' understandability, dimensionality, construct validity and reliability. Our adapted SUS statements were mostly understandable. A Principal Component Analysis resulted in a four-Component model; two of those components were established as reliable. However, we were only able to support construct validity for four questionnaire statements (and none of the four Components). This pilot study contributes to the knowledgebase of user testing with children.

Author Keywords
System Usability Scale; Mobile; Testing with Children.

CSS Concepts
• **Human-centered computing~Human computer interaction (HCI)~HCI design and evaluation methods**; *Usability Testing*;
• **Social and professional topics~User characteristics~Age**; Children.

Introduction

In this paper, we present a pilot study of our methods that included 30 children in the evaluation of the usability and learnability of mobile apps. While it is likely that children under 12-years of age have used a mobile device in the U.S. [8], there is a lack of information to guide interaction designers and usability experts when designing and testing mobile apps for/with children. Usability experts and designers concerned with creating technologies for children need to consider their differences from adults, which include cognitive, motor, social, emotional and communication abilities [4]. In the larger project (which this pilot study is part) we aim to create child-centric tools/guidelines to scaffold involving children in mobile design.

In a recent related project that supports the need for design guidelines, Soni et al. [9] conducted a literature review that resulted in a framework of recommendations for touchscreen interactions for children. While our goals are similar, our planned methods involve children as user testers. As such, we recognized that involving children as testers required us to first experiment with methods that would lead to reliable and valid findings. We began our experimentation of methods by adapting and testing the modified System Usability Scales (SUS).

SUS questionnaires use a 5-point Likert scale to assess the level of agreement to ten statements related to usability and learnability [10]. The SUS is considered a quick, reliable and valid method for assessing learnability and usability [10]. In our review, we were not able to find any adapted SUS questionnaires for children; this paper contributes, therefore, to the knowledgebase of user testing with children.

Design and testing of technologies with/for children

Several researchers have explored methods to include children in design and testing of technologies. Druin [3] established one of the most cited taxonomies of methods. She organized her taxonomy by defining a spectrum of four roles children have played in the design and testing of technologies: (1) user, (2) tester, (3) informant and (4) design partner. While the boundaries between the roles are often blurred, the essential message of Druin's taxonomy, and similar categorizations, emphasize the importance of including children in the design and testing of technologies aimed for their use. In the larger project, we plan to create tools that catalog and make actionable effective child-centric mobile design patterns; the envisioned tool will support and encourage involvement of children as design partners. In this pilot, however, our child participants acted in the role of 'testers.'

Multiple researchers have developed methods and tools for involving children as testers (e.g., [5]). In an example of exploring user testing tools for young children, Read et al. [11] created a 'Fun Toolkit' for measuring enjoyment. Researchers have also focused on comparing user testing methods with children. For example, Baauw and Markonpoulos [1] compared using think aloud and post-study interviews with 25 children between the ages of 9-11; they found that the think-aloud method identified more usability problems than the interviews. In work that used methods similar to those we piloted in this study, Budiu and Neilson [2] examined the usability of websites with a wide range of children (aged 3-12) to establish web design guidelines. Building on their methodology, we tested our methods with lab studies that included 30 children.

Age:	7-8	9-11	Total
Female	4	8	n = 12
Male	7	11	n = 18
Total	11	19	n = 30

Table 1: Child participant demographics

	Reading	Math
Grade 2	n = 5	n = 5
Grade 3	n = 5	n = 5
Grade 4	n = 6	n = 6
Grade 5	n = 5	n = 5
Grade 6	n = 9	n = 9

Table 2: Child participant grade levels for reading and math

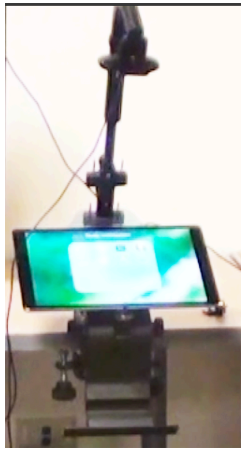

Figure 1: iPad set-up in lab

Pilot study

We designed this pilot study to answer methods-related research questions. First, (RQ1) can the SUS be reasonably adapted for children of reading age? For this pilot study, we targeted two groups of differing reading ages: 7-8 and 9-11. Second (RQ2), how will our adaptations to the SUS perform with users? We focused on RQ2 in this extended abstract paper.

Methods

This project was approved by the Institution Review Board at DePaul University.

Participants

We began with interviews of four teachers who had experience with our targeted age groups. Teachers were recruited through acquaintances. They had between 4-16 years of teaching in public elementary schools in Illinois and Wisconsin; two had experience with the younger group (7-8 years) and two with the older group (9-11 years).

Thirty child participants were recruited through a snowball method that began with recruitment postings that included a link to a screening survey; the recruitment ads were posted on multiple local message boards that catered to parents. To qualify for the study children had to: (a) be between the ages of 7-11 in June-September of 2019; (b) have at least 10 hours of experience using iPads; and (c) expected to be placed in reading and math at their grade level or no more than one year above or below. The mean age was 9.47 years (SD 1.44); the median grade level was fourth grade. See Tables 1 and 2 for additional demographics.

Data Collection

We conducted the teacher interviews in June 2019. We began by asking about their background and teaching experience. After discussing their experiences about how technologies were chosen, used and introduced in their classrooms, we reviewed the adult SUS statements for suggested re-writes. (In this paper we only report findings associated with the SUS discussion). The interviews took about an hour; participants were given a $50 gift card for participation.

Interaction with the children took place in our campus lab located in downtown Chicago. The lab is divided into activity and observation rooms separated by a one-way mirror. After obtaining parental permission and touring the lab, we asked the child participants for their assent. Parents were given a choice as to where to observe; all but one watched from the observation room. We used a 12.9-inch iPad Pro secured on a mobile mount that recorded the children's hands (see Figure 1). We also video recorded the sessions on a camera that faced the children's backs for later transcription.

We chose eight apps that were focused on teaching children coding and computational thinking; we limited the app genre and levels the children played in this pilot study so that we would have comparable design patterns to assess the effectiveness of our exploratory methods. Our criteria were that the apps needed at least a 4-star rating, a purchasable version to avoid ads and in-app purchases, and were designed for our targeted age groups. Four were designed for children 6-8 and four designed for children 9-11.

We started the sessions with a demonstration of think-aloud protocol by assembling a Lego figure, then asked

Statement 1: I think that I would like to use this system frequently = *Grade level 6*

9-11: If I had this [app] on my iPad, I think that I would like to play it a lot = *Grade level 3*

7-8: I would like to play [app] a lot more = *Grade level 1*

Statement 2: I found the system unnecessarily complex = *Grade level 11*

9-11: I was confused many times when I was playing [app] = G*rade level 4*

7-8: [app] was hard to play = G*rade level 2*

Statement 3: I thought the system was easy to use = Grade *level 2*

9-11 and **7-8:** I thought [app] was easy to use = *Grade level 2*

Statement 4: I think that I would need the support of a technical person to be able to use this system = *Grade level 7*

9-11: I would need help from an adult to continue to play [app] = G*rade level 4*

7-8: I would need help to play [app] more =G*rade level 2*

the children to do the same with a figure they chose and were able to keep. We then asked them to interact with two apps that they had not used before, completing puzzles/levels we pre-selected starting with the beginning/tutorial level. Their interaction with each app was limited to 17-minutes. After interacting with the app, we asked the children about their experience using our adapted SUS questionnaires.

Parents were given a $50 digital gift card (vendor of their choice) as a gratuity and reimbursed for their transportation. We also gave the children a $25 physical App Store gift card as a thanks for their participation. Lab sessions took about 60 minutes.

Data Analysis

Our analysis in this paper is focused on the assessment of our adapted SUS questionnaire statements (RQ2). We first combined the teacher's suggestions, and then tested the reading level for each statement using the WebFX readability tools [12]; we modified the statements as needed to meet the target grade levels.

Following procedures outlined in [10], we normalized the Likert answers to a 0-4 positive scale and then multiplied the answer totals to normalize the SUS scores from 0-100. Previous research on the SUS has indicated that the set of SUS statements are multidimensional [10]; i.e., statements 4 and 10 have been found to be associated with learnability and the remaining associated with usability. We explored the dimensionality of our modified statements through a Principal Component Analysis (PCA) using a polychoric correlation matrix (because we had ordinal data) to explore underlying components. This resulted in four dimensions; see more in findings.

To assess construct validity, we determined how many puzzles the children completed. However, this was not possible for two of the apps; one because it did not have discrete puzzles to complete and the other because it crashed several times. This resulted in 46 data points (19 for ages 7-8 and 27 for ages 9-11). We then normalized the puzzle completion rate for each app on a scale of 0-10 by assigning a score of 10 to the largest number of puzzles completed.

We then tested if we could combine the age groups for more power in our analysis. We conducted t-tests to determine if there were significant differences between the two age groups and their (a) puzzle completion rate and (b) total SUS scores. Neither was significant, for puzzle completion, $t_{(44)}= -.446$, $p = .67$ and for SUS scores $t_{(58)} = 1.03$, $p = .31$. This indicated that we could combine the two age-group's data.

To explore associations with the number of puzzles completed (i.e. our idea for construct validity), we conducted Spearman rank correlations (i.e. Spearman's rho) between the puzzle level completion rate and (a) total SUS scores, (b) SUS statements individually and (c) the resulting four PCA dimensions.

Finally, because Cronbach's alpha assumes unidimensional data, we assessed internal consistency (i.e. reliability) for each of the four PCA Components individually.

Findings

Questionnaire design

The average reading level grade for the adult SUS statements was estimated at 6.9 using readability tests [12]. In the sidebar (pages 3-6) we present the original

Statement 5: I found the various functions in this system were well integrated = *Grade level 10*

9-11: I always felt like I knew what to do next when I played [app] = *Grade level 3*

7-8: I knew what to do next when I played [app] = *Grade level 2*

Statement 6: I thought there was too much inconsistency in the system = *Grade level 8*

9-11: Some of the things I had to do when playing [app] did not make sense = G*rade level 4*

7-8: Some things in [app] made no sense = G*rade level 2*

Statement 7: I would imagine that most people would learn to use this system very quickly = *Grade level 8*

9-11: I think most of my friends could learn to play [app] very quickly = G*rade level 4*

7-8: [app] would be easy for my friends to learn =G*rade level 2*

Statement 8: I felt the system was cumbersome to use = *Grade level 5*

9-11: Some of the things I had to do to play [app] were kind of weird = *Grade level 3*

7-8: To play [app] I had to do some weird things =*Grade level 2*

adult 10-SUS statements and our adapted statements with their associated reading levels. For children aged 7-8 we aimed to keep the reading level at grade 2 or lower (sometimes settled for grade 3). For the children aged 9-11 we aimed for grade level 4 or lower (sometimes settled for grade 5).

Beyond simplifying the SUS statements, our teacher participants had three additional suggestions: (1) use the app name in the statements so there would not be any confusion about what we were referring to; (2) add statements to assess enjoyment (see 11-13); and (3) create a visual representation of the Likert scale of agreement similar to that of Read et al.'s smile scale [11]; see Figure 2.

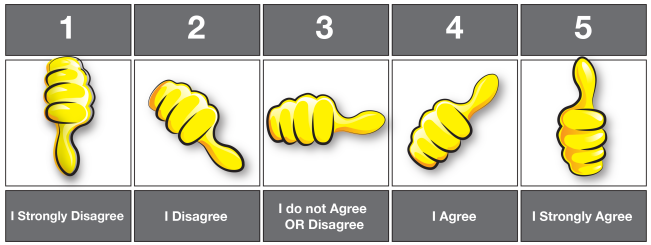

Figure 2: Visual representation of the Likert scale

Assessing Our SUS Questionnaires

We assessed the understandability, dimensionality, construct validity and reliability of the dimensions of our adapted SUS questionnaires.

UNDERSTANDABILITY

The concepts of agreement and disagreement were understandable for all child participants. Most were not confused by any of the statements; however, six children asked for clarifications on statements 6 and 8.

DIMENSIONALITY

To explore multidimensionality using a polychoric correlation matrix of the Likert answers (n = 60) we conducted a PCA with Varimax rotation. The PCA results indicated that there were four Components: (1) statements 1, 5, 9, 11, 12, 13; (2) statements 2, 3, 6 and 7; (3) statement 8; and (4) statements 4 and 10. See Figure 3. Together the factors accounted for 65% of the variance which is considered acceptable [6].

Statement	*Communalities*	*Comp 1 Loading*	*Comp 2 Loading*	*Comp 3 Loading*	*Comp 4 Loading*
1	0.636	**0.78**	0.12	-0.09	-0.08
2	0.627	0.12	**0.62**	0.42	0.23
3	0.681	0.22	**0.75**	-0.08	0.24
4	0.697	0.00	0.27	0.40	**0.69**
5	0.609	**0.55**	0.06	0.21	0.51
6	0.715	-0.10	**0.61**	0.52	-0.26
7	0.595	0.38	**0.68**	-0.01	0.14
8	0.727	0.13	0.05	**0.84**	0.06
9	0.558	**0.70**	0.25	0.05	-0.07
10	0.668	-0.27	0.14	-0.26	**0.71**
11	0.759	**0.84**	-0.05	0.23	-0.01
12	0.57	**0.71**	0.18	0.16	0.09
13	0.574	**0.63**	0.23	-0.33	-0.15
		Comp1	*Comp 2*	*Comp 3*	*Comp 4*
Unrotated Variance		*31%*	*47%*	*57%*	*65%*
Varimax Rotated Variance		*25%*	*41%*	*53%*	*65%*

Note. N=60

Figure 3: Principal Component Analysis of SUS statements

CONSTRUCT VALIDITY

We found that there was no significant correlation between the puzzle completion rate and total SUS scores using Spearman's rho, (P = .03, n = 46, *p* =

Statement 9: I felt very confident using the system = *Grade level 7*

9-11: I was confident when I was playing [app] = *Grade level 5*

7-8: I was proud of how I played [app] = *Grade level 1*

Statement 10: I needed to learn a lot of things before I could get going with this system = *Grade level 5*

9-11: I had to learn a lot of things before playing [app] well = *Grade level 4*

7-8: There was a lot to learn to play [app name] = *Grade level 2*

Statement 11 (added): 9-11: I really enjoyed playing [app]= *Grade level 5*

7-8: Playing [app name] was fun = *Grade level 3*

Statement 12 (added): 9-11 and **7-8:** If we had more time, I would keep playing [app] = *Grade level 3*

Statement 13 (added): 9-11 and **7-8:** I plan on telling my friends about [app] = *Grade level 3*

.82). We further explored Spearman correlations among the puzzle completion rate and individual statements: 2, 4, 8, and 13 were significantly correlated to the puzzle completion rate. Last, we analyzed puzzle completion for each Component.

We found using Spearman' rho, Components 1, 2 and 3 were not significantly associated with the puzzle completion rate (n = 46): for Component 1- P =-.17, p=.24; for Component 2 - P =-.18, p=.24; and for Component 3 - P =.22, p=.14. Component 4 however was very close to a significant association to the puzzle completion rate, P = .28, p = .055. (This association was largely due to statement 4's high correlation).

RELIABILITY

Finally, we conducted Cronbach's alpha to assess reliability (we could only assess this for dimensions 1, 2 and 4 because dimension 3 had only one statement). Dimensions 1(n = 60, α = .84) and 2 (n = 60, α = .71 were statistically reliable. However, dimension 4 was not (n = 60, α = .43). (An alpha between .8 and .9 is considered good and between .7 and .8 adequate.)

Discussion

We presented our experience of adapting the SUS questionnaire for two age groups of children: 7-8 and 9-11. Our modified questionnaires and our visual Likert scale were understood by our child participants with minimal clarifications. The SUS scores for each of the two groups were nearly identical indicating that our modifications were age-appropriate. Additionally, the statements in Components 1 and 2 were reliable; however, we need to consider changes to statements 6, 8 and 10 to increase understandability and reliability. While our findings were an encouraging first step, our project had many limitations that we plan to address in future work.

Limitations and Future work

We were not able to establish that our questionnaires had construct validity. This could be a deficiency in our technique; i.e., puzzle completion rate may not have been a good indication of usability. Since we were only assessing the first 17 minutes of play, we hypothesized that our puzzle completion rate may be a better indicator of learnability than it was of usability. This was supported by the close association statement 4 which has been linked to learnability [10].

Our inability to establish validity could also be linked to small sample sizes. With only 46 samples for puzzle completion, we had little power to assess relationships between SUS scores and puzzle completion leading to a potential type II error for our validity tests. With larger samples, we also would have followed our PCA with an Exploratory Factor Analysis (EFA) to test goodness of fit. EFA needs a sample size of 5-20 samples per item (we had 13 items) as a minimum [6].

Our samples were also too small for a traditional use of SUS. The sample sizes for each app in our study ranged from 5-10; the advised minimum for SUS use is 12-14 [10]. While we were not concerned in this study with the actual SUS outcomes, we will need to increase our sample sizes to confidently benchmark the usability and learnability of mobile interaction designs and design patterns in future work.

Acknowledgements

Thanks to all of our participants and to DePaul University Research Council for funding this project.

References

[1] Ester Baauw and Panos Markopoulous. 2004. A comparison of think-aloud and post-task interview for usability testing with children. In Proceedings of the 2004 conference on Interaction design and children: building a community (IDC '04). ACM, New York, NY, USA, 115-116.

[2] Raluca Budiu and Jakob Nielsen. 2010. Children (Ages 3–12 Years) on the Web (3rd Edition). Fremont, CA, USA.

[3] Alison Druin. 2002. The role of children in the design of new technology. Behavioir & Information Technology, 21(1): p. 1-25

[4] Jerry Alan Fails, Mona Leight Guha and Allison Druin. 2014. Methods and techniques for involving children in the design of new technology for children Foundations and Trends in Human-Computer Interaction. Now Publishers Inc.

[5] Andrew Large, Valerie Nesset, Jamshid eheshti and Leanne Bowler. 2007. Bonded Design: A Methodology for Designing with Children, *in Advances in Universal Web Design and Evaluation*, Sri Kurniawan, Ed. Idea Group Publishing, Hershey, PA, USA.

[6] Joseph F Hair, William C. Black, Barry J. Babin and Ralph E. Anderson. 2014. *Multivariate Data Analysis*, 7th ed. Pearson Education Limited, UK.

[7] Thomas W. Malone. 1982. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems* (CHI '82). ACM, New York, NY, USA, 63-68.

[8] Jakob Nielson *The Nielson 2016 Mobile Kids Report*. Available at https://www.nielsen.com/ Last accessed November 28, 2018.

[9] Nikita Soni, Aishat Aloba, Kristen S. Morga, Pamela J. Wisniewski, and Lisa Anthony. 2019. A Framework of Touchscreen Interaction Design Recommendations for Children (TIDRC): Characterizing the Gap between Research Evidence and Design Practice. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (IDC '19). ACM, New York, NY, USA, 419-431. DOI: https://doi.org/10.1145/3311927.3323149

[10] Jeff Sauro. 2011. *A Practical Guide to the System Usability Scale*. Measuring Usability LLC, Denver, CO, USA

[11] Janet C. Read. 2012. Evaluating artefacts with children: age and technology effects in the reporting of expected and experienced fun. In Proceedings of the 14th ACM International conference on Multimodal Interaction (ICMI '12). ACM, New York, NY, USA 241 -248.

[12] WebFX Readability Test Tool. Available at: https://www.webfx.com/tools/read-able/. Last accessed October 6, 2019.