

Think-Aloud Verbalizations for Identifying User Experience Problems: Effects of Language Proficiency with Chinese Non-Native English Speakers

Mingming Fan*

Hong Kong University of Science and Technology
Hong Kong & Guangzhou, China
mingmingfan@ust.hk

Lingyun Zhu

Rochester Institute of Technology
Rochester, NY, USA
lz2035@rit.edu

ABSTRACT

Subtle patterns in users' think-aloud (TA) verbalizations (i.e., utterances) are shown to be telltale signs of user experience (UX) problems and used to build artificial intelligence (AI) models or AI-assisted tools to help UX evaluators identify UX problems automatically or semi-automatically. Despite the potential of such verbalization patterns, they were uncovered with native English speakers. As most people who speak English are non-native speakers, it is important to investigate whether similar patterns exist in non-native English speakers' TA verbalizations. As a first step to answer this question, we conducted think-aloud usability testing with Chinese non-native English speakers and native English speakers using three common TA protocols. We compared their verbalizations and UX problems that they encountered to understand the effects of language and TA protocols. Our findings show that both language groups had similar amounts and proportions of verbalization categories, encountered similar problems, and had similar verbalization patterns that indicate UX problems. Furthermore, TA protocols did not significantly affect the correlations between verbalizations and problems. Based on the findings, we present three design implications for UX practitioners and the design of AI-assisted analysis tools.

CCS CONCEPTS

• **Human-centered computing** → **Usability testing; Empirical studies in HCI.**

KEYWORDS

language proficiency, think aloud protocols, verbalization, Chinese non-native English speakers

ACM Reference Format:

Mingming Fan and Lingyun Zhu. 2021. Think-Aloud Verbalizations for Identifying User Experience Problems: Effects of Language Proficiency with Chinese Non-Native English Speakers. In *The Ninth International Symposium*

*corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Chinese CHI 2021, October 16–17, 2021, Online, Hong Kong

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8695-1/21/10...\$15.00

<https://doi.org/10.1145/3490355.3490358>

of *Chinese CHI (Chinese CHI 2021)*, October 16–17, 2021, Online, Hong Kong. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3490355.3490358>

1 INTRODUCTION

Think-aloud protocols (TAs) are widely used in usability testing to identify user experience (UX) problems [18, 34]. During TA usability testing, participants verbalize what they are thinking while working on the task with the test interface. Their verbalizations (i.e., utterances) provide access to their invisible thought processes, which are otherwise inaccessible to UX evaluators. Despite the value of conducting TA usability testing, analyzing recorded TA sessions is often arduous and time-consuming [18, 34]. Traditional analysis methods are largely manual, which entail playing session recordings, listening to users' TA verbalizations, and observing other behavioral signals simultaneously to pinpoint UX problems. As it becomes increasingly easier to conduct a large amount of TA usability test sessions remotely via online platforms (e.g., [22, 50]), it is imperative to explore ways to improve traditional manual analysis methods.

Toward this goal, researchers recently studied users' TA verbalizations (i.e., what users say) and their speech features (i.e., how they say it) [9, 13, 20, 29, 51] and uncovered a series of subtle verbalization and speech patterns that are telltale signs of UX problems [17, 20]. For example, when users encounter UX problems, they tend to verbalize utterances of their observations and remarks than other types of utterances (e.g., action description) [17, 20]. Such TA verbalization patterns have been utilized to build artificial intelligence (AI) models to detect UX problems automatically [16]. Moreover, these patterns have also been leveraged to build human-AI collaborative analysis tools to better detect UX problems by combining the advantages of both AI and UX domain experts [19].

Despite the great potentials of TA verbalization patterns for automatically or semi-automatically detecting UX problems [16, 19], these patterns were uncovered with *native* English speakers [17, 20]. Compared to native English speakers (379 million) [11, 15], more people around the world speak English as a second language (753 million) (i.e., *non-native* English speakers) [11, 15]. Take the US as an example, almost half of the residents in America's largest cities speak a language other than English at home [21]. Over 151,000 employees from all over the world working in the US every year do not necessarily speak English as their first language [10, 47]. Consequently, it is not uncommon that non-native English speakers would participate in usability testing and be asked to think aloud in English. Thus, it is important to understand whether verbalization patterns discovered among native English speakers [17, 20] still

exist among non-native English speakers and whether there are any differences. Motivated by this problem, in this research, we took a first step to explore the following research question (RQ):

- **RQ1:** *How do English language groups (i.e., native and non-native speakers) affect TA verbalizations and UX problems?*

Furthermore, three types of TA protocols are commonly used in usability testing [3, 17, 34]: 1) Ericsson and Simon’s classic think-aloud protocol (CTA) [14], which was used to uncover the subtle verbalization patterns indicative of UX problems among native English speakers [17, 20], 2) the speech-communication protocol (SC) [3], and 3) the interactive think-aloud protocol (ITA) [12, 46]. While participants are only reminded to “keep talking” in CTA, they receive speech tokens (e.g., “Em hmm”) from the moderator in SC or are constantly probed to answer questions (e.g., “what are you looking for?”) from the moderator in ITA. In other words, participants experience different amounts of interventions while thinking aloud in these protocols. However, it remains unknown how TA protocols affect non-native and native English speakers’ verbalizations and UX problems they experience. In this research, we took an initial step to explore:

- **RQ2:** *How do TA protocols (i.e., CTA, SC, and ITA) affect two English language groups’ verbalizations and UX problems?*

To answer two RQs, we conducted online think-aloud usability testing with 18 non-native and native English speakers. As non-native English speakers of different cultures might have different thinking and speaking behaviors, we focused on a subgroup of non-native English speakers—Chinese students who studied in US universities in this research as a first step to explore this problem space. Chinese students in US universities have taken a lion share of all international students enrolled in the past decades [27] and speak English as a second language regularly in their study and daily life.

During the study, participants of two language groups (i.e., non-native and native speakers) worked on tasks with three representative websites while thinking aloud using three TA protocols (i.e., CTA, SC, ITA) respectively. We transcribed their verbalizations (i.e., utterances), categorized them into five categories following prior studies [9, 13, 17, 20], identified UX problems that they encountered, and analyzed how different verbalization categories indicate UX problems.

Our results show that non-native English speakers’ verbalizations were similar to those of native English speakers in terms of the relative proportions of different verbalization categories and the correlations between verbalization categories and UX problems. Furthermore, the trends between verbalization categories and UX problems were mostly consistent across three types of TA protocols. Based on the findings, we further discuss the implications for building AI models and human-AI collaborative UX data analysis tools. In sum, we make the following contributions:

- An initial understanding of how language groups affect TA verbalizations and their correlations with UX problems;
- An initial understanding of how three types of TA protocols affect users’ verbalizations and their correlations with UX problems.

2 BACKGROUND AND RELATED WORK

Our work was inspired and informed by related work in three areas: *Types of Concurrent Think-Aloud Protocols*, *Language Proficiency in Think-Aloud Studies*, and *Users’ Verbalizations in Think-Aloud Studies*.

2.1 Three Types of Concurrent Think-Aloud Protocols

When using concurrent think-aloud (TA) protocols, participants verbalize their thought processes while working on the tasks at the same time. Depending on the types of prompts or interventions administered by the study moderator, there are three common types of concurrent TA protocols: the *Classic Think-Aloud protocol (CTA)*, the *Speech Communication protocol (SC)*, and the *Interactive Think-Aloud protocol (ITA)*.

Classic Think-Aloud (CTA): CTA was established as a valid approach to studying human thinking processes by Ericsson and Simon [14] and later introduced into the fields of HCI and UX to study UX problems. Ericsson and Simon proposed a set of guidelines for conducting CTA: have a practice think-aloud session before the actual study session; use neutral instructions that do not direct participants to verbalize a specific type of thought process; keep prompts and interventions to a minimum by only reminding participants to “keep talking” if they fall into silence for a period.

Speech Communication protocol (SC): Realizing the unnaturalness of thinking aloud and helping to promote thinking aloud in usability testing, Boren and Ramey found that CTA can be hard to execute by UX evaluators in practice [3] and proposed the SC protocol, which asked the moderator to play an active listener role by using tokens, such as “Em hmm”, “And now? ...”, in addition to “keep talking”, to help participants think aloud.

Interactive Think-Aloud (ITA): In practice, it is not uncommon that the moderator actively asks participants questions during think-aloud usability test sessions to inquire about their opinions, explanations, or suggestions [12, 31, 46]. This variation of the concurrent TA protocols was often referred to as interactive think-aloud (ITA).

Since prior studies showed the pros and cons of these three TA protocols (e.g., [1, 42, 43]), we included all the three TA protocols to understand whether these TA protocols affect native and non-native English speakers’ verbalizations.

2.2 Language Proficiency in Think-Aloud Studies

When conducting think-aloud studies in English, prior research often recruited participants of native or fluent speakers to minimize the potential affects of language proficiency. We conducted a literature review and identified 16 papers that studied think-aloud protocols [1, 2, 5, 6, 9, 13, 16, 17, 25, 29, 30, 33, 35, 41, 49, 51]. Eleven of these papers did not specify the language proficiency of their participants [2, 5, 6, 9, 13, 25, 30, 35, 41, 49, 51]. The five papers left mentioned that their participants were either native speakers or competent in the language that they spoke in their studies [1, 16, 17, 29, 33]. To the best of our knowledge, no studies have explicitly compared think-aloud verbalizations and UX problems of native and non-native speakers of a language. In this research, we sought to fill in this literature gap.

2.3 Users' Verbalizations in Think-Aloud Studies

To understand what participants verbalize in think-aloud sessions, researchers coded participants' verbalizations and identified different **verbalization categories**. In an early work, Bowers et al. identified five verbalization categories: Procedure, Explanation, Reading, Design, and Others [4]. Later, Cooke conducted a study using a website and identified five similar categories: *Procedure, Reading, Observation, Explanation, and Others* [9]. Procedure refers to participants' verbalizations that describe "participants' current or future actions"; Reading refers to participants' verbalizations when they read any information (e.g., link labels, phrases, or sentences) from the test product; Observation refers to participants' verbalizations when they make remarks or observations about the test product or about themselves; Explanation refers to participants' verbalizations when they explain their behaviors with the test product; Others refer to verbalizations that do not fit in the above four categories. These five categories of verbalizations were later confirmed in Elling et al.'s study, in which participants used more websites than Cooke's study [13]. Although later studies divided verbalizations into more categories (e.g., [29, 51]), these categories could be mapped into the five-category scheme proposed in Cooke's study [9]. Recently, Fan et al. also adopted Cooke's five-category scheme to study the correlations between verbalization categories and UX problems among native English speakers [17]. Following these prior studies, in this study, we also adopted Cooke's categorization strategy to analyze our participants' verbalizations and investigated how these verbalization categories indicate UX problems for both non-native and native English speaking groups.

3 METHOD

We present the details of the IRB-approved user study in the section.

3.1 Participants

We recruited 18 participants through social media platforms, word-of-mouth, and snowball sampling. All participants were undergraduate or graduate students in US universities except one who recently graduated. One group of participants (N=9) was native English speakers from the US (8) and Canada (1), and the other group of participants (N=9) was Chinese non-native English speakers who studied in US universities.

Table 1: None-native English speakers' basic demographic information

ID	Education level	English Test Score	Proficiency Level
1	Graduate	TOEFL 91	C1
2	Graduate	TOEFL 96	B2
3	PhD	TOEFL 80	B1
4	Graduate	IELTS 6.5	B2
5	Graduate	TOEFL 94	B1
6	Undergraduate	TOEFL 110	B2
7	Undergraduate	N/A	C1
8	Undergraduate	N/A	C2
9	Graduate	TOEFL 106	C1

Table 1 shows the demographic information of the non-native English speakers. In addition to the standard English test scores (e.g., TOEFL, IELTS), the researchers informally assessed participants' proficiency level based on their conversations with these participants using the Common European Frame of Reference for Languages (CEFR) guidelines [40]. The definitions of the related levels were as follows: B1: intermediate; B2: upper intermediate; C1: effective operational proficiency; C2: proficiency.

3.2 Tasks

We chose three types of websites to increase variations of test products. One was an example of *information-rich* websites¹ (web 1); one was an example of *e-commerce* websites² (web 2); and the last one was an example of *productivity-enhancement* websites³ (web 3).

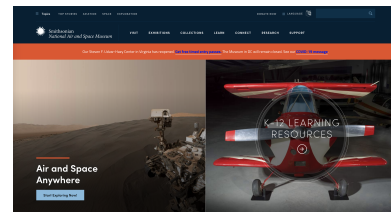


Figure 1: Web 1: an Air and Space Museum (an information-rich website)

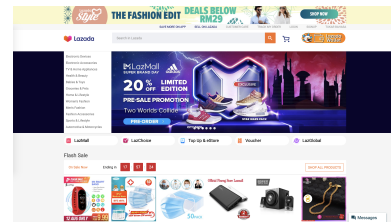


Figure 2: Web 2: Lazada (an E-commerce website)

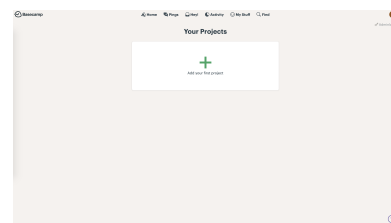


Figure 3: Web 3: Basecamp (an online teamwork tool)

Figure 1 shows the homepage of the museum website (Air and Space Museum) website. Figure 2 shows the homepage of the e-commerce website (Lazada). Figure 3 shows the homepage of the

¹<https://airandspace.si.edu/>

²<https://www.lazada.com.my/>

³<https://basecamp.com/>

teamwork tool website (Basecamp). We chose these websites because 1) they represent common types of websites one would use in their daily life; 2) we conducted heuristic evaluation using both Nielsen heuristics [37] and Norman principles [39] and found that all of these websites contained UX problems; and 3) none of the participants used these websites prior to the study.

We designed corresponding tasks that would require participants to use the features of the websites that contained UX problems. Table 2 shows the websites and the corresponding tasks used in the study.

3.3 Procedure

We conducted the user study with participants online using Zoom. We screen- and audio-recorded study sessions. The study lasted for about an hour.

Upon signing the consent form, the moderator explained to participants that they would complete tasks on three websites and think aloud during the process. To help participants better understand how to think aloud, the moderator first explained verbally how to think aloud and then showed a think-aloud demo video provided by the Nielsen Norman group [38], which shows a user working on a website while thinking aloud at the same time. Afterward, the moderator asked participants to practice thinking aloud on a website, which was not one of the test websites, to complete a task. Next, the moderator asked participants to work on the three test websites to complete corresponding tasks in Table 2 while thinking aloud at the same time. For each website, participants used one of the three TA protocols (i.e., CTA, SC, ITA) as explained in Sec 2.1. The order of TA protocols and the test websites were counterbalanced using the Latin square design in Table 3.

When conducting CTA sessions, the moderator followed Ericsson and Simon’s guidelines and did not use any prompts and only reminded participants to “keep talking” if they fell into silence for more than 10 seconds. When conducting SC sessions, the moderator followed the guidelines put forward by Boren and Ramey [3] and played an active listener’s role by saying phrases such as “Em, hmm”, “uh-huh”, “and now?”, and “keep talking” to encourage participants to think aloud. When conducting ITA sessions, the moderator actively probed participants using five types of prompts that were derived from prior studies [1, 31, 51]: *Clarifying intentions* (“What are you looking for?”), *Seeking explanations* (“Could you tell me why you did that?”), *Seeking opinions* (“What do you think of it?”), *Seeking suggestions* (“What redesign do you suggest?”), and *Seeking user expectation* (“what do you expect to be there?”).

After participants completed each task, they were asked to fill in the NASA Task Load Index (TLX) form to measure their perceived task load of completing the task while thinking aloud.

4 ANALYSES

4.1 Categorizing Think-aloud (TA) Verbalizations

We first used an automatic transcribing tool, Otter.ai [44], to transcribe session recordings and then manually checked the transcriptions to correct errors. After this process, we had all participants’ think-aloud verbalizations.

We followed a similar process used in the literature [9, 13, 17, 51] to review each think-aloud session recording and segment it into smaller segments based on pauses in the participant’s verbalizations and the semantics of the verbalizations. A segment could include sentences, phrases, or single words.

Next, for each segment, two UX researchers independently reviewed the corresponding verbalizations and assigned it a *verbalization category* label using Cooke’s five-category scheme [9] as explained in Sec 2.3. This scheme was widely adopted by prior studies [13, 17, 20, 29]. The five verbalization categories were: *Procedure*, *Reading*, *Observation*, *Explanation*, and *Others*. Table 4 shows the categories, their definitions, and examples from our participants’ think-aloud verbalizations.

After each researcher finished assigning category labels for each segment independently, they reviewed their category labels together to gain a consensus on the labels. If there was a disagreement, they explained their rationales for their labels, discussed with each other, and consolidated the category labels.

4.2 Identifying UX Problems

For each verbalization segment, two UX researchers followed the same procedure of assigning verbalization category labels in Sec 4.1 to determine whether the user encountered a problem and to assign a binary problem label (0: no problem; 1: problem). Each segment with a problem label “1” represents “a moment in which a user encountered a problem” and was referred to as a “**problem encounter**.” Because different “problem encounters” might be caused by the same underlying UX problem, two UX researchers further reviewed these “problem encounters” and combined the problem encounters caused by the same underlying problem, which was referred to as an “**actual problem**”. As a result, the number of “actual problems” would be less or equal than the number of “problem encounters.”

Furthermore, two researchers followed the same procedure to assess the severity of each “actual problem” using Nielsen’s guidelines [36]. The definitions of the five severity levels are as follows: level 0 means “no usability problem”; level 1 means “cosmetic problem”; level 2 means “minor usability problem that should be given low priority”; level 3 means “major usability problem that should be given high priority”; level 4 means “usability catastrophe that imperative to fix before product can be released” [36].

5 RESULTS

5.1 Verbalization Categories

5.1.1 Verbalization Categories grouped by Language Groups. Table 5 shows the number and percentage of segments in each verbalization category for native and non-native English speaking participants. Results suggest that the general trends for the two language groups were similar. Specifically, two most frequently verbalized categories (i.e., Procedure and Observation) and three least frequently verbalized categories (i.e., Others, Reading, and Explanation) were the same for both native and non-native English speaking participants.

One difference was that, for native English speakers, Observation was the most frequently verbalized category followed by Procedure. For non-native English speakers, Procedure was the most

Table 2: Tasks for participants to complete on each website

Websites	Tasks
Web 1 (Museum, an information-rich website)	1) Find out how much it will cost to enter the museum, and days of closure; 2) Find out what things you can do in the museum; 3) Find out what exhibitions are on view currently; 4) Find out where the exhibition (Boeing Milestones of Flight Hall) is on the museum’s map, and find out what you will see in this exhibition; 5) Find out if there is an audio guide in the museum.
Web 2 (Lazada, an e-commerce website)	1) Find out a smartwatch that meets all the 3 requirements: can be used underwater; can track your heart rate; has great customer ratings; 2) Find out what other customers said good about the watch you just found; 3) Find out where you can ask the seller questions; 4) Find out other products from the same seller; 5) Find out where you can know the shipping status of your order;
Web 3 (Basecamp, an on-line teamwork tool)	1) Create a project named “birthday party”; 2) Invite two persons into the project; 3) Share the idea of decoration you like (a link) with them; 4) Assign tasks to each person; 5) Delete the project.

Table 3: The study design with three TA protocols (CTA, SC, ITA) and test websites counter-balanced for two language groups.

Non-Native speakers’ IDs	Native speakers’ IDs	TA Protocols and websites		
P1	P10	CTA (web 1)	SC (web 2)	ITA (web 3)
P2	P11	SC (web 1)	ITA (web 2)	CTA (web 3)
P3	P12	ITA (web 1)	CTA (web 2)	SC (web 3)
P4	P13	CTA (web 2)	SC (web 3)	ITA (web 1)
P5	P14	SC (web 2)	ITA(web 3)	CTA (web 1)
P6	P15	ITA (web 2)	CTA (web 3)	SC (web 1)
P7	P16	CTA (web 3)	SC (web 1)	ITA (web 2)
P8	P17	SC (web 3)	ITA(web 1)	CTA (web 2)
P9	P18	ITA (web 3)	CTA (web 1)	SC (web 2)

Table 4: Verbalization categories, definitions and examples from the study.

Categories	Definitions	Examples
Procedure	Describe their current or future actions	"I'll start with sports and lifestyle." "Where should I go?"
Reading	Read any information (e.g., link labels, phrases, or sentences) from the test product	"The all-in-one toolkit for working remotely."
Observation	Make an observation or a remark about the test product or themselves (e.g., comments, feelings)	"Looks like it's selling everything under the sun." "So I think that makes it a little bit more overwhelming."
Explanation	Explain their behaviors on the test product	"I think underwater was probably a bad term, maybe I had to search for waterproof."
Others	Verbalizations that do not fit in the above four categories: task-related (e.g., read task descriptions or ask questions about tasks); verbal fillers (e.g. Um, Ah, alright)	"I want to buy a smartwatch to track my swimming exercise [task]." (note that this participant was paraphrasing the task) "Alright." "Let's see."

frequently verbalized category, followed by Observation. In other words, compared to native speakers, non-native speakers verbalized a relatively higher percentage of what they were doing (i.e., Procedure) than what they were remarking (i.e., Observation)

5.1.2 Verbalization Categories grouped by TA Protocols. Table 6 shows the number and ratio of verbalization segments in each category by TA protocols. For all TA protocols, Observation and Procedure were the most frequently appeared two categories, followed by Others, Reading, and Explanation.

While CTA and SC had slightly higher percentages of Procedure than Observation, ITA had a slightly higher percentage of

Observation than Procedure. In other words, with the CTA or SC protocol, participants tended to verbalize what they were doing (i.e., Procedure) more often than to make remarks (i.e., Observation). In contrast, under the ITA protocol, participants tended to make remarks (i.e., Observation) more often than verbalizing what they were doing (i.e. Procedure).

5.1.3 Effects of Language Groups and TA Protocols. To further understand the effects of language groups and TA protocols, we performed three-way ANOVA with both TA protocols and verbalization categories as within-subjects factors and language groups as the between-subjects factor. Results show 1) no significant difference

Table 5: Number (percentage) of verbalization segments in each verbalization category for each language group.

Category	Native	Non-native
Observation	752 (34.2%)	758 (33.5%)
Procedure	619 (28.1%)	793 (35.0%)
Others	434 (19.7%)	336 (14.8%)
Reading	286 (13.0%)	207 (9.1%)
Explanation	111 (5.0%)	172 (7.6%)
Total	2202 (100.0%)	2266 (100.0%)

Table 6: Number (percentage) of verbalization segments in each verbalization category for each TA protocol.

Category	CTA	SC	ITA
Procedure	508 (34.6%)	416 (33.2%)	488 (27.9%)
Observation	460 (31.4%)	395 (31.5%)	655 (37.5%)
Others	267 (18.2%)	242 (19.3%)	261 (14.9%)
Reading	168 (11.5%)	126 (10.0%)	199 (11.4%)
Explanation	64 (4.4%)	75 (6.0%)	144 (8.2%)
Total	1467 (100.0%)	1254 (100.0%)	1747 (100.0%)

for the *language groups* ($F(1, 16) = 0.062, p = 0.806, \eta_p^2 = 0.004$); 2) significant differences for *TA protocols* ($F(2, 32) = 4.621, p = 0.017, \eta_p^2 = 0.223$) and *verbalization categories* ($F(4, 64) = 44.440, p < 0.0001, \eta_p^2 = 0.71$). We further performed Scheffe Post-hoc analysis for *verbalization categories* and *TA protocols*.

For the *native language group*, Scheffe Post-hoc analysis found significant differences for the following pairs: (Explanation, Observation), (Explanation, Procedure), (Explanation, Others), (Reading, Observation), and (Reading, Procedure). Similarly, For the *non-native language group*, Scheffe Post-hoc analysis found significant differences for the following pairs: (Explanation, Observation), (Explanation, Procedure), (Others, Observation), (Others, Procedure), (Reading, Observation), and (Reading, Procedure). In other words, the significant differences were between the least and most frequently appeared categories for both native and non-native language groups.

For *TA protocols*, Scheffe Post-hoc analysis showed that they did not have any significant effect on all verbalization categories except the Explanation category. Specifically, ITA had significantly more Explanation than CTA and SC. In other words, participants tended to explain their behaviors more often in ITA than in CTA or SC.

5.2 UX Problems

5.2.1 UX Problems and Examples. The test websites were used as vehicles to answer our RQs, which focused on understanding non-native and native English speakers' think-aloud verbalizations and the correlations between the verbalizations and UX problems. Nonetheless, we present example UX problems, the usability heuristics violated [37], and participants' think-aloud verbalizations in Table 7 to better contextualize the results presented in the rest of Sec. 5.

5.2.2 UX Problems grouped by Language Groups. Table 8 shows the number of problem encounters and actual problems for two language groups respectively. To reiterate, the number of actual problems was 15 and 18 for native and non-native English participants respectively.

We further analyzed *common* actual problems in both language groups and the *unique* actual problems to each language group. Among the 19 actual problems, 14 were identified as common problems for both language groups and five were unique problems to each group. Four out of the five unique problems were of the lowest severity level 1 (i.e., cosmetic problem [36]) and the rest one was also of low severity level 2 (i.e., minor problem [36]). In other words, both native and non-native speakers' think-aloud verbalizations were equally effective in identifying UX problems of high severity levels with only minor differences in revealing UX problems of low severity levels.

5.2.3 UX Problems grouped by TA Protocols. We also counted the number of problems encounters in each TA protocol. Table 9 shows the number of problem encounters and actual problems identified for each TA protocol.

5.2.4 Effects of Language Groups and TA Protocols. For the number of problem encounters, ANOVA analysis found no significant effect of the *language groups* ($F(1, 16) = 0.252, p = 0.618, \eta_p^2 = 0.02$). This suggests native and non-native English speakers' verbalizations do not have significant difference in identifying UX problems. While ANOVA analysis found a significant effect of *TA protocols* ($F(2, 32) = 5.122, p = 0.034, \eta_p^2 = 0.24$), Scheffe post-hoc analysis did not find significant difference among any pairs of TA protocols. This suggests that the three types of TA protocols do not have significant difference in identifying UX problems.

Similarly, for the number of actual problems, ANOVA results also found no significant effect of *language groups* ($F(1, 16) = 0.313, p = 0.583, \eta_p^2 = 0.019$) or *TA protocols* ($F(2, 32) = 2.582, p = 0.091, \eta_p^2 = 0.138$). This again suggests there are no significant differences in identifying UX problems between language groups or between TA protocols.

5.3 Correlations between Verbalization Categories and UX Problems

To understand how each verbalization category is indicative of UX problems, we counted the number of segments in each *verbalization category* (i.e., Procedure, Reading, Observation, Explanation, and Others) that were associated with a UX problem. We then grouped them by *language groups* and *TA protocols*. Table 10 and Table 11 showed the results of the number of segments in each verbalization category indicating UX problems for language groups and for three TA protocols respectively.

5.3.1 Correlations organized by Language Groups. As shown in Table 10, the number of segments of each verbalization category related to a UX problem was similar for two language groups: 97 verbalization segments related to a problem in the native language group and 101 in the non-native language group. Moreover, the order of the verbalization categories indicating UX problems (from the most to the least) was also the same for both native and non-native language groups. These results suggest that language groups

Table 7: The UX problems with usability heuristics violated and the example think-aloud verbalizations.

UX Problems (<i>Usability heuristics violated</i> [37])	Problem description with think-aloud verbalizations
The design did not speak users' language or failed to match users' mental model. (<i>Match between system and the real world</i>)	Web 2 (Lazada): The options were not organized in a natural and logical order. It took P2 a while to find the target function, and she verbalized, "Where's my function? I feel like there are too many (options) here."
The design failed to provide users with effective error messages that could indicate problems and suggest solutions (<i>Help users recognize, diagnose, and recover from errors</i>)	Web 1 (Museum): The navigation disappeared when P12 tried to move the mouse to the sub-navigation. He verbalized, "Here, visit, Oops...[navigation bar disappears], visit."
The design failed to keep users informed about what is going on through appropriate feedback. (<i>Visibility of system status</i>)	Web 1 (Museum): P18 was confused about the map and verbalized, "I don't know where exactly in the map is it." Web 2 (Lazada): The workspace icon was not self-explainable and needed additional explanation. P13 verbalized, "Okay, so pretty blank here. Doesn't really tell you what you can do."
The design failed to prevent problems from happening in the first place. (<i>Error prevention</i>)	Web 3 (Basecamp): The accent color mislead users to make mistakes. Instead of sending the project out as required, P4 accidentally saved the project as a draft because the "Draft" button was green and the "Post" button was white. He verbalized, "Send as a draft... Oh, No. Post this."

Table 8: Number of problem encounters and actual problems for each language group.

Group	Problem encounters	Actual problems
Native	35	15
Non-native	41	18
Total	76	19

Table 9: Number of problem encounters and actual problems for each TA protocol.

Group	Problem encounters	Actual problems
CTA	25	12
SC	15	12
ITA	36	15
Total	76	19

Table 10: Number (percentage) of segments in each category indicating UX problems for each language group.

Category	Native	Non-native
Observation	57 (58.8%)	66 (65.3%)
Procedure	14 (14.4%)	22 (21.8%)
Explanation	14 (14.4%)	10 (9.9%)
Others	12 (12.4%)	3 (3.0%)
Reading	0 (0.0%)	0 (0.0%)
Total	97 (100.0%)	101 (100.0%)

do not significantly affect how verbalization categories indicate UX problems.

Table 11: Number (percentage) of segments in each category indicating UX problems for each TA protocol.

Category	CTA	SC	ITA
Observation	31 (55.4%)	31 (77.5%)	61 (59.8%)
Procedure	13 (23.2%)	5 (12.5%)	18 (17.6%)
Others	9 (16.1%)	3 (7.5%)	3 (2.9%)
Explanation	3 (5.4%)	1 (2.5%)	20 (19.6%)
Reading	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total	56 (100.0%)	40 (100.0%)	102 (100.0%)

5.3.2 Correlations organized by TA Protocols. As shown in Table 11, the order of the verbalization categories from the most to the least indicative of UX problems was the same for CTA and SC. While ITA followed a similar trend as CTA and SC, Explanation in ITA was more indicative of problems than in CTA or SC. Moreover, while CTA and SC had a similar number of segments related to a problem, ITA had more segments related to problems than CTA or SC.

5.4 Task Load

Table 12: Average ratings of the six scales of the NASA Task Load Index (TLX) grouped by language groups.

TLX scales (range: 1–21)	Native	Non-native
Mental demand	4.2	5.4
Physical demand	3.3	4.0
Temporal demand	4.3	5.6
Performance	15.1	16.0
Effort	4.8	5.6
Frustration	4.1	3.2

5.4.1 Task Load by language groups. Table 12 shows the average ratings of the six scales of the NASA Task Load Index (TLX) grouped by language groups. The higher the score, the higher the task load in that scale, except for the performance scale. Results show that: 1) overall, both native and non-native English speaking participants felt that thinking aloud was not much demanding and their performance was high; 2) Compared to native English speaking participants, non-native English speaking participants felt that thinking aloud was relatively more mentally, physically, and temporally demanding, more effortful, and more frustrated. However, the differences in each scale between the two language groups were not significant based on ANOVA.

Table 13: Average ratings of the six scales of the NASA Task Load Index (TLX) grouped by three TA protocols.

TLX scales (range: 1–21)	CTA	SC	ITA
Mental demand	4.8	4.9	4.8
Physical demand	3.6	3.4	3.9
Temporal demand	4.9	4.7	5.4
Performance	15.3	16.4	14.9
Effort	4.7	5.1	5.8
Frustration	3.5	3.1	4.3

5.4.2 Task Load by TA Protocols. Table 13 shows the average ratings of the six scales of the NASA Task Load Index (TLX) grouped by three TA protocols. Results show that: 1) regardless of TA protocols, participants felt that thinking aloud was not much demanding and their performance was high; 2) the differences in all scales between three protocols were not significant based on ANOVA.

6 DISCUSSION

6.1 The Effect of Language Groups

We discuss the effect of language groups on three aspects: *the proportions of verbalization categories, UX problems, and the correlations between verbalization categories and UX problems.*

Proportions of Verbalization Categories: Our analysis showed no significant difference in the proportion of each verbalization category for two language groups. Moreover, the proportions of five categories also followed a similar trend for two language groups. Observation and Procedure were the most frequently appeared two categories, followed by Explanation, Others, and Reading.

While there was no statistical significance, the non-native language group did seem to verbalize a relatively higher proportion of Procedure (i.e., what they were doing) and a relatively lower proportion of Reading (i.e., what they saw) than the native language group.

Because both Procedure and Reading were all level-1 and level-2 types of verbalizations according to Ericsson and Simon’s definition [14], we combined these two categories and found that the proportion of the two categories together was similar for the two language groups: 41.1% and 44.1% for native and non-native language groups. To better compare our results with prior work, we further combined the verbalizations of each category across all our participants. Table 14 shows the result. It showed that our

study found similar proportions of Procedure and Reading, Observation, and Explanation as Elling et al. [13], Zhao et al. [52], and Fan et al. [17], but there was a difference between ours and Cooke’s study [9] or Fan et al.’s study [20]. The difference might be due to differences in the test products, TA protocols used, and the participants. While Cooke only used CTA in their study [9], we used three types of TA protocols. While we tested with websites only, Fan et al. [20] tested with both physical and digital products. Furthermore, while Fan et al. [20] focused on older adults, our participants were all young adults. Future work should conduct more-controlled studies with the same set of products, participants, and study procedures to better understand how test products and participants’ age and other backgrounds might affect think-aloud verbalizations.

Table 14: Proportions of verbalization segments in each study. N/A means the category was not used in the study. Zhao et al.’s research [52] used two TA protocols similar to CTA and ITA in this study.

Studies	Procedure Reading	Observation	Explanation	Others
Cooke [9]	77%	10%	5%	8%
Elling et al. [13]	40%	34%	7%	19%
Zhao et al. CTA [52]	70.3%	20.1%	9.6%	N/A
Zhao et al. ITA [52]	49.9%	33.8%	16.3%	N/A
Fan et al. [17]	56.3%	37.6%	5.9%	N/A
Fan et al. [20]	31.2%	62.5%	3.4%	2.9%
Our current study	42.6%	33.6%	6.3%	17.2%

UX Problems: Our analysis results show that the language groups did not significantly affect either the number of problem encounters or the number of actual problems. The only difference was in the identification of some low severity UX problems. The implication is that both native and non-native English participants are equally effective in helping locate common and severe UX problems in think-aloud usability testing.

Correlations: Results show that correlations between verbalization categories and UX problems followed similar trends for both native and non-native language groups. The categories ranged from the most to the least indicative of UX problems were: Observation, Procedure, Explanation, Others, and Reading.

6.2 The Effect of TA Protocols

Similarly, we discuss the effect of three TA protocols on the same three aspects: the proportions of verbalization categories, UX problems, and the correlations between verbalization categories and UX problems.

Proportions of Verbalization Categories: TA protocols did not have a significant effect on all categories except Explanation. ITA had significantly more proportions of Explanation than CTA and SC. This difference was likely because the moderator prompted the participants to verbalize more by asking for explanations, opinions, and suggestions when using the ITA protocol. In contrast, such prompting behavior was forbidden in both CTA and SC protocols.

UX Problems: TA protocols also did not have a significant effect on the number of actual problems. While TA protocols were shown to have a significant effect on the number of problem encounters,

post-hoc analysis did not find a significant difference among TA protocols. This suggests that all TA protocols were equally effective in identifying UX problems.

Although there was no statistical significance, ITA found a higher number of problem encounters than CTA and SC, according to Table 9. To better understand this difference, we examined the severity of the problems found by each TA protocol. Four out of the five problems that were only found by ITS were all of the lowest severity level (i.e., cosmetic problems [36]). Cosmetic problems were mostly related to non-essential and nice-to-have features, such as changing the cursor icon when hovering over a clickable element. In other words, all protocols were equally effective in identifying severe UX problems. Furthermore, this finding was also consistent with previous research, which also found CTA, SC, and ITA found a similar number and type of problems [1].

Correlations: Correlations between Verbalization Categories and UX Problems followed similar trends for three types of TA protocols. The Observation category was the most indicative of problems of all categories. This was consistent with prior findings on CTA [17]. One difference among the TA protocols was that Explanation in ITA seemed to be more indicative of problems than in CTA or SC. This difference was likely because ITA had a higher proportion of verbalizations in the Explanation category than CTA or SC.

6.3 Design Implications

In this research, we took a first step to uncover similarities and differences in think-aloud verbalizations, the UX problems, and their correlations between *two English language groups* (i.e., native and non-native English speakers) in *three think-aloud protocols* (i.e., CTA, SC, ITA). Based on the findings, we discuss three design implications (DIs).

DI1: Non-native English speakers can be as effective as native English speakers to help identify UX problems in Think-Aloud Usability Testing. Our findings show that the verbalization patterns (i.e., verbalization categories and how they indicate UX problems) observed in native English speakers in prior research [9, 13, 17, 20, 51] were mostly applicable to non-native English speakers whose English proficiency was on or above the intermediate level [40]. Thus, UX practitioners could consider to enroll English speakers to their think-aloud usability testing for identifying UX problems without needing to worry about whether they are native or non-native speakers as long as their English reaches an intermediate level [40].

DI2: Three concurrent think-aloud protocols (i.e., CTA, SC, and ITA) are equally effective in identifying severe UX problems. Our results suggest that the three types of think-aloud protocols (i.e., CTA, SC, and ITA) do not significantly affect the number of UX problems. While ITA might be able to identify slightly more problems, these extra problems are often of low severity levels. In contrast, ITA requires significantly more effort from the study moderator. The moderator has to constantly probe participants with different types of questions, which increases her work load. As a result, we suggest UX practitioners stick with Ericsson and Simon’s classic think-aloud protocol (CTA) [14], which not only minimizes the effort from the moderator, who only needs to remind

participants to “keep talking” if they fall into silence, but also is able to uncover a similar set of important UX problems as the other two TA protocols (i.e., SC, ITA).

DI3: Non-native English speakers’ verbalizations could be used to increase the amount of training data to build artificial intelligence (AI) models to detect UX problems automatically. Our results show that similar verbalization patterns indicating UX problems were found in both non-native English speakers and native English speakers. For example, as indicated in Sec 5.3, for both language groups, the most verbalized categories were Observation and Procedure, which were also more indicative of UX problems than other categories. Such patterns could be used to build AIs to help UX practitioners better identify UX problems. For example, Fan et al. showed that verbalization category label was an effective feature to train AIs to detect UX problems automatically along with other textual and acoustic features, such as sentiment, speech rate, pitch and loudness [16]. Meanwhile, verbalization category information could also be visualized to direct UX evaluators’ attention to segments of a TA usability test video that are more likely associated with a UX problem [19].

One key challenge in building such AIs or AI-assisted tools is to gather a large amount of training data. This is even more challenging if UX practitioners have to focus on recruiting native English speakers due to the concerns of the effect of participants’ language proficiency on identifying UX problems. Indeed, outside of a few countries where English is the first language, it is almost impossible to recruit a sufficiently large number of native English speakers. Luckily, many countries around the world have a critical mass of regular non-native English speakers. Our research provides evidence that non-native English speakers’ TA verbalizations can be as effective as those of native speakers for identifying UX problems. Thus, the implication is that when UX practitioners recruit participants to collect a large amount of TA usability test videos, such as via online remote usability testing (e.g., [2, 49]), to train AI models, they could consider to recruit both native and non-native English speakers. This would significantly increase their chance and lower their cost of getting a sufficient amount of TA test sessions to train AIs.

6.4 Limitations and Future Work

Our research contributes to a first understanding of the effects of English language proficiency on think-aloud verbalizations and how TA verbalizations of native and non-native English speakers indicate UX problems. In this section, we would like highlight the limitations of our current work and discuss potential future work.

First, our study included a relatively small number of participants ($N = 18$). It is essential to replicate the study with more participants of diverse backgrounds to further validate and extend the findings.

Second, we chose a subgroup of non-native English speakers—Chinese students who studied in US universities. Other non-native English speakers, such as ones from other countries, may think aloud differently in English. Further, our non-native English speakers had intermediate or higher language proficiency (Table 1). Non-native English speakers with lower proficiency might verbalize their

thoughts differently too. Thus, it is worth investigating how cultural backgrounds and language proficiency of non-native English speakers might affect their thinking aloud in English.

Third, we focused on English language. Other language speakers might organize their thoughts differently. Slobin argued that language “is a subjective orientation to the world of human experience, and this orientation affects the ways in which we think while we are speaking” [48]. Indeed, research suggested that cultures and languages could affect users’ think-aloud processes [7, 8, 26]. Thus, it is worth exploring whether subtle verbalization patterns are telltale signs of UX problems for other languages, for example, Chinese, Dutch and French to name a few. If similar verbalization patterns indeed exist in other languages, then such patterns could be utilized to inform the design of AIs or AI-assisted tools to support UX evaluators to uncover UX problems for products used by people who speak those languages.

Fourth, conducting TA usability testing relies on the moderator’s skills too, especially the interactive think aloud (ITA). Our study moderator had two years of UX experience in conducting TA sessions. It is possible that the moderator’s experience would affect how she prompts participants (e.g., what prompts to use and how frequently to prompt participants). Thus, it is worth exploring how the moderator’s experience might affect how participants think aloud.

Fifth, we selected three types of websites to increase the variations among the test products. However, it remains unknown whether and how products might affect participants’ thinking aloud processes. Similarly, task difficulty might influence how participants think aloud too. For example, participants might find it harder to verbalize their thoughts when working on a more difficult and demanding task. Thus, more research is warranted to understand how products and task difficulty might affect non-native English speakers’ verbalizations and their correlations with UX problems.

Lastly, we used three types of *concurrent* think-aloud protocols in the study. To reduce the potential drawbacks of concurrent protocols, *retrospective* think-aloud protocols (RTAs) are used in usability testing [18]. When using RTAs, participants complete the task first and then verbalize their thought processes while watching their session recording. Verbalizations in RTAs are shown to be a valid representative of participants’ thoughts [24]. It is worth investigating how participants’ think-aloud verbalizations indicate UX problems in RTAs in the future.

7 CONCLUSION

Recent studies showed that subtle patterns in TA verbalizations are telltale signs of UX problems. However, such studies were conducted with native English speakers [17, 20]. There are more non-native English speakers around the world, who might think and verbalize thoughts differently than native speakers due to different cultural backgrounds. In this research, we took a first step to explore this problem space by studying a subgroup of non-native English speakers—Chinese students who study in US universities. We have compared non-native English speakers’ verbalizations in three common types of concurrent TA protocols (i.e., the classic TA protocol [14], the speech-communication TA protocol [3], and the interactive TA protocol [12, 31, 46]) with those of native English

speakers. Our findings show that for both non-native and native English participants, their verbalization categories and the relative proportions, the UX problems that they encountered, and the correlations between the verbalization categories and the UX problems were largely similar. Moreover, the findings were mostly consistent for three TA protocols.

Analyzing TA test session recordings to identify UX problems often entails reviewing video recordings and listening to users’ think-aloud verbalizations. This process is often arduous and time-consuming, which has motivated researchers to explore ways to automate or semi-automate this analysis process [16, 19]. Such methods leveraged subtle patterns in users’ think-aloud verbalizations and speech patterns [17, 20]. Our findings support that subtle verbalization patterns uncovered in previous studies with native English speakers [17, 20] are largely applicable to non-native English speakers. One implication is that UX practitioners could recruit both native and non-native English speakers to participate in TA usability testing to gather a larger amount of data, which could be used to train AI models to detect UX problems automatically or semi-automatically [16, 19, 23, 28, 32, 45]. As an initial exploration of this problem space, we only studied one subgroup of non-native English speakers. As culture can affect thinking and speaking behaviors, future work should investigate TA verbalizations of different subgroups of non-native English speakers and of different languages to better inform the design of automatic- or semi-automatic analysis methods for identifying UX problems that users of different cultures and languages might encounter.

REFERENCES

- [1] Obead Alhadreti and Pam Mayhew. 2017. To intervene or not to intervene: an investigation of three think-aloud protocols in usability testing. *Journal of Usability Studies* 12, 3 (2017), 111–132.
- [2] Morten Sieker Andreasen, Henrik Villemann Nielsen, Simon Ormholt Schröder, and Jan Stage. 2007. What happened to remote usability testing? An empirical study of three methods. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1405–1414.
- [3] Ted Boren and Judith Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication* 43, 3 (2000), 261–278.
- [4] Victoria A. Bowers. 1990. Concurrent versus retrospective verbal protocol for comparing window usability. *Doctoral Dissertations* (1990).
- [5] Anders Bruun, Peter Gull, Lene Hofmeister, and Jan Stage. 2009. Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1619–1628.
- [6] Kapil Khalil Madathil and Joel Greenstein. 2011. Synchronous remote usability testing - A new approach facilitated by virtual worlds. *Conference on Human Factors in Computing Systems - Proceedings*, 2225–2234. <https://doi.org/10.1145/1978942.1979267>
- [7] Torkil Clemmensen. 2011. Templates for cross-cultural and culturally specific usability testing: results from field studies and ethnographic interviewing in three countries. *Intl. Journal of Human-Computer Interaction* 27, 7 (2011), 634–669.
- [8] Torkil Clemmensen, Morten Hertzum, Kasper Hornbæk, Qingxin Shi, and Pradeep Yammiyavar. 2009. Cultural cognition in usability evaluation. *Interacting with computers* 21, 3 (2009), 212–220.
- [9] Lynne Cooke. 2010. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication* 53, 3 (2010), 202–215.
- [10] American Immigration Council. 2020. The H-1B Visa Program: A Primer on the Program and Its Impact on Jobs, Wages, and the Economy. <https://www.americanimmigrationcouncil.org/research/h1b-visa-program-fact-sheet>.
- [11] David Crystal. 2003. *English as a global language*. Ernst Klett Sprachen.
- [12] Joseph S Dumas, Joseph S Dumas, and Janice Redish. 1999. *A practical guide to usability testing*. Intellect books.
- [13] Sanne Elling, Leo Lentz, and Menno De Jong. 2012. Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions on Professional Communication* 55, 3 (2012), 206–220.

- [14] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data*. the MIT Press.
- [15] Ethnologue. 2020. What is the most spoken language? <https://www.ethnologue.com/guides/most-spoken-languages>.
- [16] Mingming Fan, Yue Li, and Khai N Truong. 2020. Automatic Detection of Usability Problem Encounters in Think-aloud Sessions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2, Article 16 (May 2020), 24 pages. <https://doi.org/10.1145/3385732>
- [17] Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Truong. 2019. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5, Article 28 (July 2019), 35 pages. <https://doi.org/10.1145/3325281>
- [18] Mingming Fan, Serina Shi, and Khai N Truong. 2020. Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *Journal of Usability Studies* 15, 2 (2020), 85–102.
- [19] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N Truong. 2020. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26, 1 (2020), 343–352. <https://doi.org/10.1109/TVCG.2019.2934797>
- [20] Mingming Fan, Qiwen Zhao, and Vinita Tibdewal. 2021. Older Adults' Think-Aloud Verbalizations and Speech Features for Identifying User Experience Problems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 358, 13 pages. <https://doi.org/10.1145/3411764.3445680>
- [21] Center for Immigration Studies. 2018. Almost Half Speak a Foreign Language in America's Largest Cities. <https://cis.org/Report/Almost-Half-Speak-Foreign-Language-Americas-Largest-Cities>. (Accessed on 08/28/2020).
- [22] FullStory. 2021. FullStory | Robust Analytics, Session Replay, Heatmaps, Dev Tools. <https://www.fullstory.com>.
- [23] Julián Grigera, Alejandra Garrido, José Matías Rivero, and Gustavo Rossi. 2017. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies* 97 (2017), 129–148.
- [24] Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. 2006. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1253–1262.
- [25] Maaikje Haak, Menno De Jong, and Peter Schellens. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers* 16 (09 2004). <https://doi.org/10.1016/j.intcom.2004.07.007>
- [26] Marinda Hall, Menno De Jong, and Michael Steehouder. 2004. Cultural differences and usability evaluation: Individualistic and collectivistic participants compared. *Technical communication* 51, 4 (2004), 489–503.
- [27] Melanie Hanson. 2021. International Student Population & Enrollment Statistics [2021]. <https://educationdata.org/international-student-enrollment-statistics>.
- [28] Patrick Harms. 2019. Automated usability evaluation of virtual reality applications. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–36.
- [29] Morten Hertzum, Pia Borlund, and Kristina B Kristoffersen. 2015. What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction* 31, 9 (2015), 557–570.
- [30] Morten Hertzum, Kristin D Hansen, and Hans HK Andersen. 2009. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology* 28, 2 (2009), 165–181.
- [31] Morten Hertzum and Kristina Bonde Kristoffersen. 2018. What do usability test moderators say? 'mm hm', 'uh-huh', and beyond. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. 364–375.
- [32] JongWook Jeong, NeungHoe Kim, and Hoh Peter In. 2020. Detecting usability problems in mobile applications on the basis of dissimilarity in user behavior. *International Journal of Human-Computer Studies* 139 (2020), 102364.
- [33] Emiel Kraemer and Nicole Ummelen. 2004. Thinking About Thinking Aloud: A Comparison of Two Verbal Protocols for Usability Testing. *Professional Communication, IEEE Transactions on* 47 (07 2004), 105–117. <https://doi.org/10.1109/TPC.2004.828205>
- [34] Sharon McDonald, Helen M Edwards, and Tingting Zhao. 2012. Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication* 55, 1 (2012), 2–19.
- [35] Sharon McDonald and Helen Petrie. 2013. The effect of global instructions on think-aloud testing. *Conference on Human Factors in Computing Systems - Proceedings*, 2941–2944. <https://doi.org/10.1145/2470654.2481407>
- [36] Jakob Nielsen. 1994. Severity Ratings for Usability Problems. <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>.
- [37] Jakob Nielsen. 1994. *Usability engineering*. Elsevier.
- [38] NNGroup. 2014. *Demonstrate Thinking Aloud by Showing Users a Video*. <https://www.nngroup.com/articles/thinking-aloud-demo-video/>
- [39] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
- [40] The Council of Europe. 2021. Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). <https://www.coe.int/en/web/common-european-framework-reference-languages>.
- [41] Erica Olmsted-Hawala, Elizabeth Murphy, Sam Hawala, and Kathleen Ashenfelter. 2010. Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. *Conference on Human Factors in Computing Systems - Proceedings* 4, 2381–2390. <https://doi.org/10.1145/1753326.1753685>
- [42] Erica L Olmsted-Hawala, Elizabeth D Murphy, Sam Hawala, and Kathleen T Ashenfelter. 2010. Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2381–2390.
- [43] Erica L Olmsted-Hawala, Elizabeth D Murphy, Sam Hawala, and Kathleen T Ashenfelter. 2010. Think-aloud protocols: Analyzing three different think-aloud protocols with counts of verbalized frustrations in a usability study of an information-rich Web site. In *2010 IEEE International Professional Communication Conference*. IEEE, 60–66.
- [44] Otter.ai. 2020. Otter Voice Meeting Notes - Otter.ai. <https://otter.ai/>.
- [45] Fabio Paternò, Antonio Giovanni Schiavone, and Antonio Conti. 2017. Customizable automatic detection of bad usability smells in mobile accessed web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.
- [46] Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: how to plan, design and conduct effective tests*. John Wiley & Sons.
- [47] Congressional Research Service. 2020. H-2A and H-2B Temporary Worker Visas: Policy and Related Issues. <https://fas.org/sgp/crs/homsec/R44849.pdf>.
- [48] Dan Slobin. 1996. *From "thought and language" to "thinking for speaking"*. Re-thinking linguistic relativity. Cambridge: Cambridge University Press.
- [49] Katherine E Thompson, Evelyn P Rozanski, and Anne R Haake. 2004. Here, there, anywhere: Remote usability testing that works. In *Proceedings of the 5th conference on Information technology education*. 132–137.
- [50] UserTesting. 2021. UserTesting: The Human Insight Platform. <https://www.usertesting.com/>.
- [51] Tingting Zhao and Sharon McDonald. 2010. Keep talking: an analysis of participant utterances gathered using two concurrent think-aloud methods. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. 581–590.
- [52] Tingting Zhao, Sharon McDonald, and Helen M Edwards. 2014. The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behaviour & Information Technology* 33, 2 (2014), 163–183.